Education Policy Analysis Archives

Volume 6 Number 10 May 22, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal. Editor: Gene V Glass Glass@ASU.EDU. College of Education Arizona State University,Tempe AZ 85287-2411

Educational Standards and the Problem of Error

Noel Wilson
School of Education
The Flinders University of South Australia

Abstract

This study is about the categorisation of people in educational settings. It is clearly positioned from the perspective of the person categorised, and is particularly concerned with the violations involved when the error components of such categorisations are made invisible.

Such categorisations are important. The study establishes the centrality of the measurement of educational standards to the production and control of the individual in society, and indicates the destabilising effect of doubts about the accuracy of such categorisations.

Educational measurement is based on the notion of error, yet both the literature and practice of educational assessment trivialises that error. The study examines in detail how this trivialisation and obfuscation is accomplished.

In particular the notion of validity is examined and is seen to be an advocacy for the examiner, for authority. The notion of invalidity has therefore been reconceptualised in a way that enables epistemological and ontological slides, and other contradictions and confusions to be highlighted, so that more genuine estimates of categorisation error might be specified.

Contents

Acknowledgments

I wish to acknowledge the help of staff and students at the Flinders Instituite for the Study of Teaching for their help, support, encouragement, stimulation and companionship over the past three years.

In particular I want to acknowledge the support of my supervisor, John Smyth, for his courage for accepting me as a student in the first place, for his clear and incisive help when I asked for it, and sometimes when I didn't, and most importantly for showing me that there are still persons working in hierarchical systems who have been able to maintain their integrity in their search for truth and justice.

About the Author

Noel Wilson is an ex-teacher, researcher, writer who has now officially retired and lives in the Adelaide Hills in South Australia. He still writes stories and novels which search in vain for publishers. He is deliighted that his mind works better now at seventy than it did at forty. Every now and then he has a little foray back into the educational field. He is a long odds optimist because he believes that sooner or later schools will get better. And he'd be pleased to engage in dialogue about this thesis. For more specifics about the author, read Chapter 2.

He can be reached at noelwilson26@hotmail.com.

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Part 1: Positioning

Chapter 1: Positioning the study: content and methodology

Chapter 2: Positioning the writer: experience

Chapter 3: Positioning the writer: philosophy and value

Chapter 1: Positioning the study - content and methodology

Summary of the study

The project grew out of a general critique of assessment theory and practices, and in particular of the way in which the notion of error in measurement is obfuscated.

The fundamental research question that informed this study is:

> How is error in measurement of standards obscured in most practical events involving assessment of persons?

The study that subsequently developed

- Clearly positions the writer in terms of the experience, philosophy and values that he brings to this study.
- Develops some tools of analysis of the educational assessment process that enables a more stringent critique of the nature and extent of error in the measurement of standards.
- Establishes the centrality of the notion of the educational standard to the categorisation, production and control of the individual in society.
- Shows how the professional literature on educational measurement is based on the notion of error, and at the same time trivialises that notion.
- Re-examines some of the fundamental assumptions of educational assessment generally and psychometrics in particular. Indicates some of their most blatant self-contradictions and fudges.
- Reconceptualises the notion of invalidity, and positions the field of educational categorisation here, from the perspective of the examined, rather than with validity, which is an advocacy for the examiner.
- Applies some of this analysis to a study of competency standards in general, and in particular University grades, and national literacy testing as developed in the Australian context during the 1990s.

As can be seen, the initial research question has generated action as well as understanding, a tool to repair the damage resulting from the critique, and a way to reduce some of the violence it implies.

Relevant Literature

The relevant literature is extensive as well as intensive, as the Bibliography shows. The extensiveness was necessary, as many of the misconceptions and fudges and contradictions that characterise the field of educational assessment have been caused by a myopia regarding knowledge outside the arbitrary boundaries within which the field encloses itself.

Within the field of educational measurement the critical studies which most overlap mine are: in the United Kingdom, Hartog & Rhodes (1936), Cox (1965); in the United States, Hoffman (1964), Nairn (1980), Airasian (1979), and Glass (1978); in Australia, Rechter & Wilson (1968).

The Hartog & Rhodes study clearly showed the enormous instability of the measurement of standards in Public Examinations in England. The sneakiness of some of the research techniques in no way detracts from the dramatic incisiveness of the data. Cox did a similar job and ended up with a similar horror story on measurements of University grades. Hoffman directed his critical attention to the detail of multiple choice testing. Nairn's critique of the work of Educational Testing Service, and in particular the part it plays in College Entrance, is devastating in its implications. Airasian's book is a comprehensive critique of competency testing. Glass attacks the measurement of standards at its most vulnerable point; there are no standards, or at least none that psychometrics can produce. And Rechter & Wilson's study indicates the confusion about how to reduce error that accompanies public examining in Australia.

On the other hand, most of the literature on reliability and validity is pertinent to this study, because, when its discourse is repositioned from examiner to examined, it provides more than enough invalidity information to self destruct.

Most studies of error in the measurement of standards are however much more specific in their focus than is mine. Their minimal effect on practice has perhaps partially been due to the fact that their critiques were in terms of their own discipline of educational measurement; a discipline that owes its very existence to the claim to accurate judgments. In terms of general style and scope this study is perhaps closer to the work of Persig (1975; 1991), who delved, articulately if deviously, much more deeply into the notion of quality.

Within the field of power relations and the construction of the individual the studies most similar are those published in Foucault and Education (Ball,1990), in particular those that take off from Foucault's placement of the examination as a central apparatus of power/knowledge.

This study is significant in that it brings these two diverse fields of educational assessment, and the power relations that pervade education, into much closer contact, to expose their interrelations, and allow the critique to cross fertilise.

Importance of the study

The initial question addressed is how the whole matter of error in measurement of standards is obscured in most practical events involving assessment and measurement.

This is directly related to the centrality of the notion of the educational standard to the categorisation, production and control of the individual in society. For if the notion of the standard is crucial to the maintenance of power relations, and its empirical realisation is prone to enormous error, then the whole apparatus of power/knowledge that depends on it is in jeopardy.

I argue in Chapters 4 and 5 that the examination normalises and individualises, and is impotent without the notion of the measured standard, the sword that divides, the wedge that produces the gaps; and how important it is that these measures of standards be seen as accurate if current societal structures are to be maintained.

One view of immorality is that it is behaviour that destabilises a social system. So if playing the game is inevitable, is questioning the rules not so much dangerous as despicable, immoral to the point of being unthinkable? Is this the reason for the great silence about the enormous errors in any measure of standards? Does this account for the erasure from public consciousness and discourse of the obvious fact that educational standards as a thin accurate line have no empirical existence, and attempts to measure in relation to that line no instrumental reality?

In Chapters 6 to 17 thirteen sources of invalidity that contribute to the error and confusion of all categorisations of individual persons are detailed and elucidated, indicating how this silence in professional and public consciousness might be filled with a deafening noise.

In Chapters 18 and 19 of this study I apply some of the analytic tools developed to the contemporary scene in Australia, and demonstrate how the noise may be turned into a coherent critique of practice. In 1997 competency standards, as a form of assessment, have become, and are becoming, the major credentialing instrument for both educational and vocational courses and jobs. In addition, they are now the basis for job descriptions. In defining what training is required for a job, what prerequisites are required to attempt a job, what the job is, and how performance on the job is to be assessed, the cycle of fantasy created by this controlled semantic reductionism is complete; the material world of education and employment has become textualised in terms of competencies (Collins, 1993; Cairns, 1992). The fragility of this theorising is exposed when examined in terms of the reconstructed notion of invalidity developed in this study.

In Universities students are still categorised in terms of grades loosely defined. What do they mean? How error prone are they? And in the schools all Australian states have agreed to introduce tests of literacy. Certainly they will introduce tests. But what will they measure? And with what accuracy? Again the reconstructed notion of invalidity is used to critically evaluate such questions.

Methodology and the critique of practice

The study roves beyond the artificial constraints of psychometric theory and test practice; into ontology, epistemology and the metaphysics of quality; into the nature of instrumentation; into the relations between equity and assessment frames of reference; into the fundamental notion of comparability; into the detail of the relation between rank orders, standards and categorisations; and into the minefield of the psychometric fudge.

Is there method in this diverse madness? Where is the methodology that informs this wild profusion? The study aims to expose the madness that underlies much of the current method. So what is a methodology that undermines methodologies?

One such method is critical analysis, the analysis of the educational discourse that comprises the field of assessment. The polices and practices of educational assessment become fused in the discourse in which they are embedded (Ball, 1994).

> Discourses are about what can be said, and thought, but also about who can speak, when, where and with what authority. Discourses embody the meaning and use of propositions and words. Thus, certain possibilities for thought are constructed . . . We do not speak a discourse, it speaks us. We are the subjectivities, the voices, the knowledge, the power relations that a discourse constructs and allows (p22).

Analysis of such discourses may not be used to determine the truth. Yet such analyses may be very sensitive to the uncovering of untruths, by determining the extent to which they embody "incoherencies, distortions, structured omissions and negations which in turn expose the inability of the language of ideology to produce coherent meaning" (Codd, 1988, p245).

How would such untruths be established?

- First, by uncovering self contradictions, within the overt discourse, or between the unstated assumptions of the discourse and the facts that the discourse establishes.
- Second, by exposing false claims, claims that may be shown with empirical evidence constructed within its own frame of reference to be untrue.
- Third, by detailing some of the psychometric fudges on which many assessment claims depend to maintain their established meaning.
- Fourth, by indicating how repositioning the discourse may dramatically change its truth value.
- Fifth, by establishing four discrete epistemological frames of reference for assessment discourse as currently constructed, and indicating the confusion when one frame is viewed from the perspectives of the others.
- Sixth, by noticing frame shifts within a particular discourse, with the resulting confusion of meaning.
- Seventh, by exposing the ontological slides and epistemological camouflages necessary to sustain many truth claims.

So in this study I will substantiate the contention that some of the explicit and implicit "truths" embedded in assessment practices are falsifiable; that empirical data constructed from their own assumptions denies the accuracy they assume; that this data is not only adequately detailed in the literature, but further, that the notion of error is the epistemological basis of much of that literature. All of which makes the public silence about the presence of error even more puzzling.

I shall show that the epistemological and ontological grounds for the whole field of assessment of individual persons are enormously shaky. I shall also explain how the literature about the very notion of validity is founded on a biased position, so that the sources of invalidity are much deeper and wider than is admitted in practice, even though clearly implied in theory and its attendant discourse.

I shall indicate the complexity of the notion of invalidity, with its practical face of error. Error includes all those differences in rank ordering and placement in different assessments at different times by different experts; all the confusions and varieties of meaning attached to the "construct" being assessed; and all those variabilities arising out of logical type errors, issues of context, faulty labelling, and problems associated with prediction. To further complicate the matter error has a different meaning depending on the assessment frame of reference. And I will show that estimates of the extent of the confusion along many of these dimensions may be easily estimated.

This is a critical study. Foucault (1988) says:

> There is always a little thought even in the most stupid institutions; there is always thought even in silent habits. Criticism is a matter of flashing out that thought and trying to change it: to show that things are not as self-evident as one believed, to see that what is accepted as self-evident will be no longer accepted as such. Practising criticism is a matter of making facile gestures difficult (p155).

Using Foucault's terminology, this is a critical study designed to make facile assessment gestures about standards difficult.

Methodology and inquiry systems

After a twenty three page discussion on data and analysis relevant to construct validation, which to Messick (1989) means all validation, he concludes

> . . . test validation in essence is scientific inquiry into score meaning - nothing more, but also nothing less. All of the existing techniques of scientific inquiry, as well as those newly emerging, are fair game for developing convergent and discriminant arguments to buttress the construct interpretation of test scores (p56).

I would broaden this to refer to any categorisation produced by transforming a continuity into a dichotomy. And for now I want to leave aside the obvious bias in the word "buttress," and focus here on inquiry systems themselves. For Messick (1989), conservative as he is, accepts that

> because observations and meanings are differentially theory-laden and theories are differentially value-laden, appeals to multiple perspectives on meaning and values are needed to illuminate latent assumptions and action implications in the measurement of constructs (p32).

Churchman (1971), elucidates five such scientific inquiry systems of differential values and epistemology, roughly related to philosophies espoused by Liebniz, Lock, Kant, Hegel and Singer. Mitroff (1973) has developed and summarised Churchman's systems. Very briefly, the Liebnizian inquiry mode begins with undefined ideas and rules of operation, ending with models that count as explanations. The Lockean mode begins with undefined experiential elements, and uses consensual agreement to establish facts. The Kantian system shows the interdependence of the Liebnizian and Lockean modes, and uses somewhat complementary Liebnizian models to interrogate the same Lockian data bank, to ultimately arrive at the best model. The Hegelian mode uses antithetical models to explain the same data, leaving it for the decision maker to create the most appropriate synthesis for a particular purpose. In this mode values of enquirer and decision maker become exposed. Finally, the inquiry system of Singer (1959), is one of multiple epistemological observation, where each inquiring system is observed from the assumptions of the others, and each methodology is processed by those of the others. Churchman (1971) paraphrases Singer clearly and cleanly: "the reality of an observing mind depends on it being observed, just as the reality of any aspect of the world depends upon observation" (p146).

How do these inquiry systems link to the seven ways of demonstrating untruths, or nonsense, detailed in the previous section? It is the Singerian inquiry mode that best characterises this study as a whole. Although particular modes have been utilised for particular critical purposes, this is in itself justified by the Singerian inquiry mode.

So whilst the first three methods listed are clearly in the Liebnizian and Lockean modes, the other four involve the explication of shifting sets of assumptions, and belong to the Singerian mode. In particular the examination of compatibilities between the four frames of reference for assessment on the one hand, and equity definitions, power relations, instrumentation requirements, and notions of comparabiltiy and quality on the other, demonstrate clearly that to the Singerian enquirer, "information is no longer merely scientific or technical, but also ethical as well" (Mitroff, 1973, p125).

The "conversation pieces" and "stories" used to demonstrate the absurdity of some assessment claims belong to the Hegelian mode. Churchman (1971) explains:

> The Hegelian inquirer is a storyteller, and Hegel's thesis is that the best inquiry is the inquiry that produces stories. The underlying life of a story is its drama, not its "accuracy". Drama has the logical characteristics of a flow of events in which each subsequent event partially contradicts what went before; there is nothing duller than a thoroughly consistent story. Drama is the interplay of the tragic and the comic; its blood is conviction, and its blood pressure is antagonism. It prohibits sterile classification. It is above all implicit; it uses the explicit only to emphasise the implicit (p 178).

Strategy of deterrence

The general strategy used to make the case for the invalidity of most current assessment practice is borrowed from military policies of nuclear deterrence. It is a strategy of overkill. Of the thirteen sources of invalidity developed in this study, any one would, if fully applied to current assessment practices, take them out, neutralise them, render them inoperable. To nullify this attack on validity of tests, examinations and categorisations generally, it is necessary to destroy not one missile, but all of them.

Methodology and structure of the study

The study has been presented in seven parts: Positioning, Context, Tools of Analysis, Error Analysed, Synthesis, Application, and a Concluding Statement.

Part 1 - Positioning : All descriptions of events, all writing, is positioned; makes certain assumptions, is viewed from a particular perspective. Part one positions the study in terms of focus and method, and the writer in terms of experience and philosophy.

In this opening chapter I position the work in terms of its general content and methodology, and show how it all fits together. So Chapter 1 briefly summarises what the study is about, what literature is most similar in both content and style, what is the importance of the study and its possible impact, and in this section how it is structured.

In Chapter 2 I show how the study is positioned in terms of some of the learnings accrued from the professional and life experiences of the author.

In Chapter 3 I indicate how the study is positioned in terms of philosophy and value, and how that relates to some contemporary literature.

Part 2 - Context: Assessment involves events that occur in, and are given meanings in, a social context. In Part 2 I elucidate some aspects of that context.

In Chapter 4 I focus on the way power relations both violate and produce those who act out their lives within their influence. In particular the centrality of the examination is exposed in the production of the modern individual, defined as an object positioned, classified and articulated along a limited set of linear dimensions.

In Chapter 5 the argument in Chapter 4 is applied and developed in terms of educational assessment. In particular I examine the crucial part that the standard plays in the whole mechanism of defining cut-offs for abnormality and non-acceptance, and how important it is that these standards be seen as accurate if current societal structures are to be maintained.

In Chapter 6 I focus on the cultural meanings that attach themselves to the notion of the standard, and assign the idea of the human standard to the mythological sphere, a place apart from critical thought. I examine the emotional intensity of discourse about the standard, its significance as an article of faith, and how this is related to the maintenance of control and good order.

Part 3 - Tools of analysis: In Part 3 some tools for looking at specific assessment events are developed. In Chapters 7 to 12 I examine four different epistemological frames of reference for assessment, and relate these to notions of equity, to hierarchical structures, instrumentation, comparability, rank orders and standards, logical types, and quality. These chapters introduce some independent, fundamental, and rarely discussed aspects of underlying assumptions involved in events culminating in the assessment of students. Inadequacies in any one of these aspects would, in a rational world, be enough to destroy the credibility of most student assessments. I will contend that all practical assessments of people contain major inadequacies in most of them.

In Chapter 7 four different frames of reference are defined; four different and largely incompatible sets of assumptions that underlie educational assessment processes as currently practised: First is the Judges frame, recognised by its assumption of absolute truth, its hierarchical incorporation of infallibility; second is the General frame, embedded in the notion of error, and dedicated to the pursuit of the true score; third is the Specific frame, which assumes that all educational outcomes can be described in terms of specific overt behaviours with identifiable conditions of adequacy; fourth is the Responsive frame, in which the essential subjectivity of all assessment processes is recognised, as is their relatedness to context.

Because of their contradictory assumptions, slides between frames result in confusion and compound invalidity.

Chapter 8 shows how certain assessment frames are inherently contradictory to certain definitions of equity, themselves contradictory to each other and to the power structures in which they are enmeshed. As such, those assessment frames and notions of equity that contradict the enveloping hierarchical structure will be seen, accurately and probably unconsciously, as potentially destabilising, and will consequently be ignored, nullified, or corrupted into acceptability.

Chapter 9 looks at Instrumentation. In this chapter we look at the conditions and invariances required in events involving measuring instruments if such events are to have credibility; in particular the notion of a Standard that theoretically defines the scale, and its confusion with a standard of acceptability, which is to be measured by the instrument, and which requires a scale in order to be located.

The various assessment modes are analysed in terms of their instrumental error. On these grounds alone all are found to be invalid.

Chapter 10 takes up the issue of comparability. What can be compared? Fundamental distinctions between more and less, better and worse are examined , their relations with uni and multi dimensionality shown, and the implications for rank ordering of students in tests and examinations unearthed. This leads to further examination of the differential privileging of sub groups and individuals when marks are added. The essential meaninglessness of such additions becomes apparent.

In Chapter 11 the relationship between rank order and standard is teased out in more detail: In particular the meanings given to the standard in the Judge and General frames of reference; how logical confusions proliferate when discourse jumps from one frame to the other; and how all categorisations involve standards and rank ordering, even though many advocates of "qualitative" assessment methods may want to deny this.

Chapter 12 leads from the implications of the Theory of Logical Types for assessment practices to an examination of the distinction between standard and quality. When the standard is seen, realistically, as unable to perform its function, quality is the notion with sufficient mythical, ideological, and intellectual status to replace it. This would produce a very different learning milieu.

Part 4 - Error analysed: In Part 4 the tools developed in Part 3 are used to discriminate particular sources of confusion and error within assessment events designed to categorise students.

In Chapter 13 the meaning of error in each frame of reference for interpreting assessments is considered. As the meaning of error changes with assessment mode, so do the methods designed to reduce such error. Procedures to reduce error in one frame are seen to increase it in another. From a perspective of oversight of the whole assessment field, this is another source of confusion and invalidity, particularly as it is rare for any practical assessment event to remain consistently within one frame of reference.

Chapter 14 addresses the question: What does a test measure? In terms of social consequences the answer is clear. It measures what the person with the power to pay for the test says it measures. And the person who sets the test will name the test what the person who pays for the test wants the test to be named. The person who does the test has already accepted the name of the test and the measure that the test makes by the very act of doing the test. So the mark becomes part of that person's story and with sufficient repetitions becomes true.

My own conclusion is that tests have so many independent sources of invalidity that they do not measure anything in particular, nor do they place people in any particular order of anything. But they do place them in an order, along a single line of "merit," and that is all they are required to do.

Chapter 15 shows some of the ways in which psychometricians fudge; by reducing criteria to those that can be tested; by prejudging validity by prior labelling; by appropriating definitions to statistical models; and by hiding error in individual marks and grades by displaced statistical data, and implying that estimates are true scores. A number of specific examples of fudging are detailed.

In Chapter 16 some of the more recent work on validity is discussed, and its positioning as advocacy demonstrated. I conclude that in practice the very existence of validity is established, validity is indeed made manifest, through the denseness of the arguments about invalidity criteria used to refute such existence, together with the reassurance that the battle continues, and some gains have been made.

Reliability is also discussed as a problematic, rather than as an obvious prerequisite to validity. I conclude that most of the mechanisms designed to increase reliability necessarily decrease validity.

Part 5 - Synthesis: In Chapter 17 the notion of invalidity is reconceptualised, having both discursive and measurable components. Thirteen (overlapping) sources of error are examined, all contributing to the essential invalidity of categorisations of persons.

Part 6 - Application: In Chapter 18 I apply the philosophical and conceptual positioning, tools of analysis, and the reconceptualised sources of error developed in this thesis to the competency based assessment policies and practices of Australia in the 1990s. I show how the notion of competency standards is overtly central to the whole competency movement, the introduction of which is shown to be overtly politically motivated. Thus the crucial links between political power and educational standards that are argued for in Chapters 3 and 4 become transparent. I then go on to examine the invalidity of competency standards in the light of the thirteen sources of error specified in the previous chapter.

Chapter 19 presents two specific applications of invalidity sources; the first relates to national literacy testing, and the second to University grades.

Impact

Assessment practice is permeated with mythology and ideology; with confusions and contradictions; with epistemological and ontological slides; with misrepresentations of frames of reference for different assessment modes; with logical type errors and psychometric fudging, in which the constructs that determine error--labelling, construction, stability, generality, prediction--are either ignored or severely constrained in the determination and communication of error, in those rare cases where personal error and likely miscategorisation is publicly admitted.

I have no expectations for this study, but some hopes. A whistle blowing study is like a joke--its impact is a function of timing. And the best timing can only be determined in retrospect. My hope is that it will lead to a reduction of the violence that is attributable to the suppression of error in the categorisation of people.

Chapter 2: Positioning the writer: experience

Introduction

As I take the epistemological position that all knowledge is based on experience, value and reflection, and all experience is influenced by prior knowledge, it seems important to indicate some of those life experiences that led me to the particular ontological and epistemological positions that inform this study. To do otherwise is to infer either their universal superiority, or their complete arbitrariness.

In this brief autobiographical note I outline some of those significant life experiences and concomitant learnings as they impinge on this study. This is neither arrogance nor self-indulgence (Mykhalovskiy, 1996). For if thirty years working in the field of educational research and assessment is not relevant to this project, then either the work, or the project, or both, must surely be trivial.

Education

This study has had a long gestation. Forty nine years ago I sat for my matriculation examination in English. I had a choice of four essays, and chose one called "Examinations." I rubbished them, unwisely it seems. I got a B grade which compared unfavourably with the second highest mark for English at my prestigious public school. That I'm still at it today indicates that non-conformity is not necessarily related to inconsistency or nonperserverence. What I learnt from this experience is that meaning and judgment are affected by context, and that appropriateness is one criterIon for the recognition of quality.

Two years of study in the University Engineering faculty convinced me that I did not want to be an engineer, and left me with one invaluable legacy; on every engineering drawing the measurement of each dimension, and the limits of accuracy within which the product must be fabricated, are indicated. In practice, because error was inevitable, the statement of acceptable error was as important as the magnitude of the dimension. Keeping within acceptable error was a major determinant of quality of product. This practice of indicating errors in measurement continued for calculations in Physics, the subject of one of my majors when I transferred to the Science faculty.

I decided to become a teacher. Moving to Education was a culture shock. I could only write scientific prose - sparse and unadorned, tight and dry, logical and on the surface devoid of any emotional involvement. So writing two thousand word essays was a problem; I generally said all I had to say in two hundred, and regarded the rest as superfluous padding. I could state my case, but had lost my personal voice.

What I learnt about assessment was at the level of "helpful hints to beginning teachers." The massive literature on educational assessment and evaluation was then, as it is now for most teachers, unknown to me. I was trained for survival, not for problematising tradition. I learnt what was implied. The game of testing had produced me, so it couldn't be all that bad.

Teaching

I taught in high schools and tested students more or less the way I'd been tested. Maybe a few less essays and considerably more short answer questions. The process was simple. I sat down, wrote some questions to comprise an examination paper, the students did it, I marked it, added up the marks, and then gave them a percentage or converted it to a grade. How was it done? Easy! Was it a problem? No! How accurate was it? Nobody, including me, ever asked!

After three years I joined the Royal Australian Air Force as an Education officer, teaching some basic physics to photographers, some nuclear physics to air crew, and some instructional technique to officers. Because I was teaching it, I learnt the technology of lecturing. It was assumed I could accurately assess all this. I averaged about six lectures a week, so they were very well prepared. With so much time, I diverted myself by writing pantomimes and musicals. I was beginning to find my voice.

Two years of work at the RAAF School of Technical Training had me writing syllabuses as well as teaching basic maths and physics. I talked to electrical fitters who had come back for training after two years in the field as electrical mechanics. None of them had used any of the eighty odd hours of mathematics in the Mechanics course. I suggested to the administration that they save time and money by leaving out the mathematics. It was explained to me that its relevance to work was irrelevant. It was necessary for the high level of trade classification. I was beginning to understand the economic and political character of credentialing.

Assessing

My last year in the RAAF was spent in the trade testing section. Fifty item, two hour, multiple choice tests were used to credential students who had spent from three to twelve months in training programs, with hundreds of hours of practical and theoretical assessment as part of the course. My attempts to point out the absurdity of this were usually met with the response that it didn't matter, because they just kept on doing the trade tests till they passed. I was becoming aware that in the world of work, as well as in the world of education, ritual was more important than rationality.

Teaching again

Observing that the influence Education Officers had on training seemed to diminish as they were promoted, I went back to teaching in a private coeducational high school. I found that what had taken twenty hours to teach to highly motivated technicians took five times as long to teach to supposedly more intelligent high school students. In my second year I told the matriculation physics class I did not intend to teach them. Rather I would try to create an environment in which they could learn. I would assume they could read the syllabus and the text book. They worked individually or in groups, developed their own notes, devised their own experiments. They completed the course by the end of June, after which I agreed to give some consolidating lectures, and class time was spent doing past examination papers and improving answers. That, after all, was the task on which they would be judged. Their results in the external examination were extremely high. I had learnt to separate the ritual of teaching from the facts of learning.

Next year I tried the same process. The students refused to cooperate. They collected notes from other schools. They insisted I teach them. After a month I had little choice. We went back to "normal" teaching methods. They got "normal" results at the end of the year. I learnt that dependency has as much attraction as autonomy, for the price of autonomy is personal responsibility.

Two other events were significant over this period. The first was a question asked by Michael, a student; What exactly is an electron? I had no idea. The question had never occurred to me. I'll let you know, I blustered. A month and many hours of reading later, I responded. Do you remember, Michael, you asked me what an electron is? No, he answered. I'll tell you anyway, I said, unperturbed. I wrote "Properties of an electron" on the blackboard, and under that heading listed some of them. The class looked on in silence. I looked at Michael. Yeah, he said, those are its properties, but what exactly is it? Ah, I said, now that's a question you'll have to ask the Rabbi. I had started to grapple with ontology. I was thirty years old.

Writing

The second involved the writing of A programmed course in Physics (Wilson, 1966). This was a linear program covering year 11 and 12 Physics. In reviewing what I had written I was dissatisfied with the presentation of force field theory. Finally I wrote this part as a dialogue between a physicist and a student. The result was much more satisfying in that the nature of a field in physics could be discussed as a problematic, rather then presented as a scientific conclusion. My first excursion into epistemology required discourse rather than didactic prose to communicate its meaning.

Assessing again

Because of my experience with multiple choice tests in the RAAF, I had been working with Australian Council for Educational Research on the construction of multiple choice physics tests. When a full-time position came up I applied for it. For the next six years I was to work as a test constructor. I learnt a lot about the nature and mechanics and rituals of testing, about the truisms and tricks of the trade. For example, that only "items" between thirty and seventy percent difficulty were chosen because others did not contribute economically to the separation of students; that seemingly almost identical questions often had very different difficulty levels; and it was almost impossible to tell, without prior testing, how difficult a test item was.

Central to the theme of this study, I also learnt, at the level of practice and praxis, the great secret about error, about the fallibility of the human judge, about the vagueness and arbitrariness of the standard. Not in that language, of course. Psychometrics provides a more prophylactic discourse about marker reliability and predictive validity and generalizability. Even so, it was impossible to miss the point. Or was it? I did a course in educational measurement at a local university to sharpen up my theoretical skills. We learnt the statistical theory and all the little techniques for reducing error, like short answer questions and multiple marking. And at the end of the course--a three hour essay type examination marked by the lecturer and then given a grade. And nobody said a word! Even more amazing, when I raised the matter with a few of the other students, they seemed unaware of the contradiction. I was learning that tertiary studies do not necessarily invoke reflective critical thinking.

There were two other outcomes of this experience of constructing test items that were important. The first related to the discourse, the arguments about the best answer that characterised the panel meetings. The second related to the values and effects of this particular testing program, and how to deal with that (Wilson, 1970).

As we got better at writing "distractors" for multiple choice questions, we found advocates among the "expert" panel for some of the distractors as the best answer, rather than the one chosen by the test writer. Of more potential educational significance was the argumentation itself, and its effect on our ability to think sharply and clearly within the fields being discussed. Tests themselves can never produce improvement in individual performance; but our experience suggested that argumentative discourse about test items could. A serendipidous piece of research at one school confirmed this. One hundred students thus engaged for about twenty hours raised test scores on each of three multiple choice papers by half a standard deviation, despite the ACER publications that claimed these tests could not be "taught" (Wilson, 1969).

The second experience related to educational values, and our attempts as "examiners" to grapple with this. None of the full-time test constructors approved of the Commonwealth Secondary Scholarship tests as an educational intervention. They were a politically inspired election gimmick. We were aware that they would have an influence on what schools taught, and possibly how they taught, even though they were supposed to be "curriculum free" as well as value free. As a result we took "educational value" as a major criteria for test validity, at least at the level of our own personal discourse. The material we chose for tests must face the question "would education be improved if teachers did try to prepare students for this sort of exercise, for answering these sorts of questions on these sorts of information or issues, for engaging in this sort of thinking and problem solving?" I was learning that no test was value free, and that these tests were certainly informed by a (possibly idiosyncratic) view of educational relevance.

Groups

During these years I also had my first experience in unstructured groups, and experienced at first hand the power of such group interactions to produce major changes in social behaviour in the participants; within the microcosmic society of such groups, as they developed, there was opportunity to take risks, revisit social experience, and re-construct social meanings. I learnt how powerful such groups could be in raising awareness, loosening counterproductive behaviours, and reframing experiential meanings (Slater, 1966).

Research

When at age forty I was appointed to head the newly established Research and Planning Branch in the SA Education Department, a position I held (with planning dropped half way through), for the next thirteen years, my major claim to expertise was in the area of testing and assessment. The Directors never allowed this to influence their decisions about committee membership, and during my sojourn with them I was never appointed by them to any departmental committee concerned with assessment. Nor, for that matter, am I aware of any decision made by the Department that was informed by research that the Branch carried out. When research knowledge was consistent with Departmental policy assertions it was utilised; when it didn't or wouldn't serve those interests it was ignored. I was learning that research knowledge was an instrument of power, a weapon for rationalising decisions, rather than a springboard for rational decision making (Cohen & Grant, 1975).

It was partly this insight, as well as a belief that my clients were students and teachers rather than administrators, that determined that most of my own research would be concerned with classroom practice. I also noticed that most educational research dealt with special groups and special problems, leaving the "normal" educational assumptions and practices unsullied by any critical research probes. So I directed most of my action research to the "average" classroom; that is, I sought out the commonalities of educational experience rather than the differences.

In the first few years I spent considerable time with teachers looking at improving assessment practices in schools. One thing in particular became apparent during these discussions--that most of what I had learnt as a professional test constructor was irrelevant to the assessment issues that concerned teachers in classrooms; these were not the sort of descriptions that helped children learn better, or helped teachers teach better. When I wrote Assessment in the primary school in 1972 the then Director of Primary Education wrote a foreward in which the final paragraph stated "some people would question his suggested limitation on testing. Whatever one's views, teachers will find the report thought provoking and valuable". In other words, I disagree with him, but respect his different viewpoint. As Directors became more managers and less educators in the 1980s, this sort of clarity and openness, this up front honesty, was to become increasingly rare.

Politics

In 1974 a thirteen year old schoolgirl was suspended from her high school and refused to accept the suspension on the grounds that it was unfair. She returned to the school and was subsequently removed forcibly by police. The incident resulted in a Royal Commission, and the Royal Commissioner found that the girl and her parents were a "trinity of trouble makers". (Royal Commission, 1974). It was never suggested that the setting up of the Commission had anything to do with the fact that the girl's father was an endorsed labour candidate and a personal friend of the Minister of Education, and that the Principal of the school was the brother of the shadow Minister of Education. Nor was it ever suggested that the united front of the Education Department officers and secondary principals had anything to do with the highly conflictual situation then existing between the Minister and the high school principals.

I thought that most of the overt conflict at the school was due to communication problems between the girl and certain members of staff, and certainly not due to the severity of the crime, which was trivial. In such cases it seemed to me to be the job of the professional staff, not the student, to resolve the conflict. So I gave evidence on behalf of the student. I was the only member of the Department to do so. What I learnt from this episode was that the structural violence embedded in institutions is evidenced not by the severity of the punishment when rules are breached, but by the severity of the punishment when the sanction, whatever it is, is not accepted. I could see that accepting any sanction reinstates the power structure; in fact, breaking the rule enables such re-establishment to become visible, enhancing the power relations. But not accepting the sanction is extraordinarily threatening because it destabilises the power structure, challenging its very existence. It also became clear to me that none of the Departmental officers, or the Royal Commissioner, could see this.

Social development research

As the development of social skills was a major objective in the stated curriculum of almost all school subjects, I initiated a major project on social development. It lasted four years, attracted two major grants, and at one stage involved six full time and six part time researchers (The Social Development Group, 1979). As a starter to this I took six months long service leave and a round the world trip. I spent some time visiting people and relevant projects in the United States, Canada, and England. I talked to teachers at primary, secondary, and tertiary levels about the social development of their students, and how they were able to facilitate that development. They all described the social development of their students during a year, whether six or twenty six years old, in the same terms; tentative, inarticulate, immature to confident, articulate, sensitive. It was obvious that what they were talking about had little to do with developmental skills.

My experience in unstructured groups suggested to me that it had everything to do with developing groups, with the way that power, affect and trust relations change if they are allowed to. I had already spent six months reading the literature on social skill development. It was often interesting, but utterly uninformative in regard to classroom practice. And we had asked teachers to describe mature social skills; they responded with good descriptions of conforming behaviour. I could see that shifting the focus to the social group, to the context of social action, produced an array of possible teacher interventions, informed by group development theory. We started with a project about developing social skills. We ended with a project on developing the classroom group; for only in a developed group would the demonstration of mature social skills be appropriate.

Rebelliousness

One incident that occurred on this journey deserves a mention, as it relates to the question of what constitutes experience. In London I went into a coma for two weeks, during which time I convulsed and hallucinated and was fed by a drip and lost 12 kilograms in weight. I was diagnosed as having viral encephalitis.

My hallucinations had a clear story line. They all involved adventures with semi humanoid monsters who were trying to kill me. The final scene had me lying on an operating table with ten humanoid gun barrels at my head. The odds were stacked against me, and death was immanent. I had time only for one statement. "You will only kill me," I said, "to prove that I cannot control you. Yet if you kill me for that, then I have completely determined your actions." They left, I came out of coma, and requested some food. With some trauma, I had learnt that the rebel is as tied to the system as the conformist. If I wanted to change the system, I would have to take a different stance; one of autonomous action, rather than rebellious reaction. I would need to tap the ambivalence of those in power, not their antagonism.

Back in Adelaide, the social development project got under way. I read the literature on (small) group development theory, and realised that most of the models could be reframed in terms of distributions of power and affect relations; and because of my physics background, I conceptualised these in terms of fields; properties of the space between rather than of the agents mediated by the fields. My personal ontology was developing, and ten years later more complex notions of power relations (eg Foucault) would find nourishment in my conceptual space.

Politics again

Part of the condition of the research grant was that separate reports be written for the major participants in the study; researchers, administrators and curriculum writers, teachers, students. I wrote the booklet for students. It was entitled How to make your classroom a better place to live in (The Social Development Group, 1980). It described the four stages of development of the classroom group, how students might experience these stages, and how they might respond to that experience. Four different responses to each situation were constructed, and were overtly categorised as positive and negative; the negative responses, with which students would identify and be familiar, were likely to be not constructive in moving the group onward; the other two responses, one involving individual action and one group action, were ones which might help the group develop. The booklet was designed for classroom discussion.

Before the book was distributed a question was asked in the South Australian parliament about the book. Was it not encouraging students to respond negatively? The Director General responded by ordering that the book be shredded. Flattered if furious with this treatment, I pointed out the conditions of the grant, and requested specific information about exactly what was objectionable in the book, so that it could be amended and reprinted. After some months the answer came back; two words, "fascist" and "fairy," had to be removed; the positive responses must come first; and there must be an overt statement that the positive responses were "better". In addition, only teachers involved in developing their class groups could distribute this book to their students.

I interpreted this to mean that there was nothing specifically at fault with the book. It was the ideology of the book, with its implicit aim of empowering students, that had caused the over-reaction. Yet the rhetoric about schools applauded the empowerment (autonomy) of students. Unwilling to confront the contradiction, the Department had to settle for limitation rather than complete suppression. For of course developing the classroom group meant that the power relations between teachers and students changed. If this happened in enough classrooms not only classroom structures, but school structures, would have to change. The implications of the research were radical rather than progressive.

Inservice training was essential if the findings of the research were to be propagated, if practice were to follow theory. So four researchers, now highly skilled in working with teachers, were retained for a year to produce inservice materials and work in schools with teachers. A year later, despite protestations, all had been returned to classrooms. An invaluable human resource for the dissemination of ways of developing the classroom group was annihilated. Fifteen years later teachers still struggle with rebellious classrooms and search for answers in individual psychology, curriculum statements still highlight the development of social skills rather than the social context for mature social behaviour, and teachers still say "groups don't work" because they don't understand group development theory. In 1980, I was beginning to learn what I knew by 1990; that nothing really changes unless the power structure changes, and hierarchical power structures are immensely stable and resistant to change (Wilson, 1991).

Consciousness

One further event in 1979 is pertinent to this story. At Findhorn, an intentional community in Scotland, I experienced some shifts in consciousness (without drugs or intention, with detachment and interest), that seemed very similar to those experiences described by mystics, and generally described under the rubric of the perennial philosophy. (Bucke, 1901; Huxley, 1946; Wilbur, 1977,1982,1991; Wilson, 1992). These experiences, and subsequent ones, make it impossible for me to take Freud's easy way out (Freud, 1963), and discount such events because I have not experienced them. Such experiences have been immensely significant in the history of the past three thousand years, for they have provided the bases for the world's great religions. The mythologies and structures that are the social manifestations of these initiating mystical events have taken very different cultural forms, but all have retained, within their core practices, considerable congruency with their source as a particular state of consciousness. This is important because it points to one exit from the maze of confusion created by the acceptance of the relativity and cultural determination of all human values (Wilbur, 1995).

Peace and violence

By 1982, Ronald Reagan's unique combination of monstrous stupidity and apocalyptic hardware had stirred the coals of fear still glimmering under the weight of twenty years of psychic numbing and denial, of human refusal to seriously consider the high probability of a nuclear holocaust that could destroy all life on the planet. Everywhere the peace movement flourished. Learned journals of all sorts from medicine to engineering, from physics to art, began to feature articles about nuclear war and its effects. Most unlikely bedfellows, Marxists and churchmen, pacifists and retired admirals, feminists and builders labourers, would all shout out their protests.

Where were the children in all this? I decided to find out. There was some American data from surveys. I decided to tap a richer source; children's fantasies of the future. The data was devastating (Wilson 1985). For many it was a post-nuclear war world, barren landscapes and destruction everywhere. For nearly all it was dehumanised, people existing either as passive recipients of technology, at the best comfortably mindless in a plastic world, at the worst slaves of the machines or robots that grind mercilessly along their efficient and pre-programmed paths. An unstoppable high-tech, high-destruct world.

Like many who start with a naive view of peace as the absence of war, my reading and reflection soon led to more sophisticated understandings; towards peace as the absence of fear at a psychological level, and as incompatible with injustice and repression at the social level. And I began to understand how injustice was often not so much a matter of human intention, as a product of historical man-made structures, continually reproduced through the human facility of role-taking, and the moralities and ideologies that are able to transform efficient violations into noble virtues. At fifty I was beginning to articulate a world-view.

During the international year of peace, schools were all expected to get involved. Believing that in dealing with violence we should begin in our own back yards, I prepared a kit for schools entitled Programs to reduce violence in schools (1986). It included ideas for involving students, teachers and parents, for collecting information, and for taking action at a school level. It also included a paper on understanding violence, in which I tried to make overt the links between violence, school structures, social control, and justice. Complete with words of encouragement from the Director General of Education, the kit went off to one hundred high schools in South Australia. One school got the project off the ground and collected data from students and staff. Then they stopped. During the year, many schools planted trees for peace. I was developing a feel for the absurd.

Writing again

Two years before, buttressed by a report by the head of another educational research organisation, the Department disbanded ours. I was sent out to graze in the country at Murray Bridge for two years as an Assistant Director Curriculum, where I managed to get two of the social development advisers back into business, before I retired gracefully. There was nothing further I could do within the system. I was ready to write, and had two young daughters at home that I wanted to spend more time with. I was learning the difference between jousting with windmills and hitting my head against a brick wall; one is a noble quest, the other just plain masochism.

The writing and the daughters got together into a book called With the best of intentions (Wilson, 1991). The book deals with the structural violence embedded in the hallowed institutions of family and school. I had decided to self-publish the book before I began, and as a result was able to give clear reign to my personal voice(s) and style. The book is egalitarian in that it treats children as fully human persons; it is iconoclastic in that it challenges many of the sacred myths and structures of child-rearing; it is written with passion and humour. It is informed by empirical data and overt in its philosophical world-view. The arguments are dense, but the presentation is, I hope, sufficiently varied and light to make its message accessible. With modifications that are essential to the context, I hoped to use a similar approach in this thesis.

The current study

A large number of significant learnings have emerged for me from the current study. I want to refer to the two that I have found the most significant. The first relates to my extensive reading of Michael Foucault, the second to my grapplings with ontology.

There were two major insights from Foucault; the first was his analysis of how culture produces and expresses rather than reduces and represses; that if the person is one dimensional, this is not because society has taken away the other dimensions, but that society, through its relations with the person, has produced a one dimensional person. The second insight was the centrality given to the examination, in all its forms, to the construction of the individual in the modern world. It was from this springboard that I could leap to observe the standard as the bullet in the examination gun.

An equally important learning from Foucault relates not to insight, but to style; not to his immense data base and sometimes lugubrious argumentation, but to the soaring rhetorical passion that marks his insightful conclusions; his demonstration that "scientific" writing does not need to be dull and portentous, but can legitimately use the full creative resources of the language, helped me to feel much more comfortable in using my own voice for this work.

My own philosophical gropings into what is knowable, what is describable, led to some surprising conclusions. Such delving was necessary, because any assessment is a description. In practice it is a description of a performance of some kind in context, even if in theory it purports to be a description of some attribute or quality of a person; this I had known for a long time. To move from here to the insight that all knowledge is a description of events involving a relationship between at least two elements, and thus to appreciate the slide made when the description is pinned to one particular element, represented a major reframing of much of my earlier thinking.

Summing up

There are at least five levels in all this: The events that I was a part of; the manifest behaviour that constituted my part of those events; my particular recall of that experience; the meanings I verbally constructed from that recalled experience; and the meanings and reactions that you, the reader, construct from all that.

Truth is not an issue here. Awareness and truthfulness are. I can only assert my truthful intentions. Regardless, the reader will make his or her own judgment about the value of the position from which they interpret me as coming.

Chapter 3: Positioning the writer: philosophy and value

Preview

In this chapter I spell out in more detail the philosophical stance that I take in this study, so that my assumptions about social life and social relations are up-front.

Whilst these assumptions are consistent with the learnings of the autobiographical sketch give in the last chapter, I have not felt it necessary, or advisable, to enter into any sort of justifying dialogue regarding my position. This is not a philosophical study, and I have always regarded justification as a loser's game.

So I have presented my philosophical position as a set of assertions with an internally consistent logic; I have briefly described the epistemological, ontological, and axionomic assumptions that have informed this study, and described how that position fits into current post-positivist, interpretivist, and post-modern paradigms.

The chapter ends with a brief outline of the assessment process constructed from my particular position.

Philosophical assumptions : What is knowledge? What is truth?

I will call an event any interaction where a change or a difference is observed or otherwise sensed (Bateson, 1979). Interactions involve some relation between elements of the event. Differences involve some relation between the elements, or the states of an element over time, that constitute the difference. So all events involve some relation between elements. And because all events involve a perception, so all events involve a perceiver. The perceiver may be automated as an instrument that senses the difference or reacts to or records the change. As Maturana (1987) expresses it, "Everything is said by an observer" (p65).

Any experience is experience (action, feeling, perception) of an event, either directly, or as recalled or as transformed in memory or action. So all experience involves relations. As all knowledge must finally depend on experience, all knowledge involves knowledge of relations; so all knowledge is constructed out of relational events.

To experience an event does not necessitate giving a meaning to that event, but does require a state of awareness or consciousness, from which the event is viewed. For example, an experience may be represented by a pattern or abstract painting which embodies relations without embodying meaning. Giving a meaning to an event requires some theoretical underpinning, some ideas or ideals; some knowledge of relations derived from other events, or possibly, if mathematical relations are construed to constitute meaning, derived from acts of imagination that transcend (are transformations of) known relations. Mathematics can be regarded as a special case of patterning, and whether mathematical propositions or systems have meaning in themselves is moot. I don't think they do. Some post-structuralists want to deny experience that excludes meaning and thus language. My experience denies their denial. Their assumptions refute my denial. Stalemate. But then, I'm writing this thesis.

I use the term meaning to involve more than prediction, which mathematics can sometimes help to accomplish. Meaning involves some reason, some purpose, some intention, some value. Thus meaning is inevitably embedded in language, itself embedded in human discourse. Unless we take a mystical view and define the meaning as the experience itself, or rather as a particular encompassing experience, in which case discourse stops and the world in its oneness pulsates. In this thesis I shall hold to the more mundane view. To do otherwise is not to proceed.

In this epistemology, experience precedes pattern, and pattern precedes meaning. "Whether we are talking about unicorns, quarks, infinity, or apples, our cognitive life depends on experience" (Eisner, 1990, p31). Meaning will then usually in its turn, but not necessarily, pre-empt and distort experience, which will then in its turn influence events. Buddhist meditation is designed to limit this distortion; which brings its participants on this issue close to post-positivists like Phillips (1990), who seem ultimately to define objectivity as the reduction of bias of various sorts.

Meaning is socially constructed because language is socially constructed. What passes for knowledge in common language is a social concurrence in a particular culture about acceptable meanings embedded in discourse. On the other hand, experience is constructed out of relational events not necessarily linked to any particular culture, and the construction of patterns or relations in response to that experience may also sidestep, or transcend, social patterning or common meanings. In other words, I hold the view that creation is immanent in all events, and in all perception of events, and change is more than the imposition of some random variation. Usually, however, we may assume that patterns are also culturally influenced.

Data is a particular form of knowledge constructed by particular people for particular purposes. Such purposes always involve the construction or isolation of events in which the observer is directly, or indirectly through associated theory, involved; for example, measuring devices involve the observer at one step removed. Thus all data, being knowledge, is constructed from events, constructed and/or observed for particular purposes. All data, to be used, must have either a predictable pattern, or a meaning, or both. So if data is to be useful, it must have links to other relational events, or have links to (uneventful) abstract relations.

It follows that, in this world, there are as many potential truths about an event as there are experiences of the event. To the extent that all experiences of the event are the same then there is a case for "the" truth. But how would this be known? Any attempt to know this would involve the sharing of meanings, which are certainly socially constructed and can be as varied as the cultures and relations and metaphors that are used to make sense of them and communicate them. So agreement about one meaning, one truth, represents conformity about social construction as much as it does concomitance of experience.

Ironically, in a social context the idea of multiple truths is unificatory, whilst the notion of one truth is fundamentally divisive; in practice the notion of one truth contradicts the collaborative ethic and supports interaction characterised by entrenched positions. Search for "the" truth is often productive within a closed space of cultural assumption, but does not lead to open inquiry outside that space; rather it invokes defensiveness, and if necessary violence in order to sustain its inviolability. Inevitably it leads to fragmentation and conformity, as contradictory elements break away to form their own "truthful" reality, and all else becomes subservient to "truths" current fashion (Feyerabend, 1988).

One more point about multiple truths; such a claim does not contain the inference of the catastrophic consequence that all "truths," that is, socially acceptable beliefs, are equally useful or sustainable, or that some cannot be falsified. At least at the level of physical definition, it is demonstrably false that I am constructed entirely of green cheese. Such a claim is not a valid contender for any claim to a truth beyond that of a very idiosyncratic and metaphorical form. Truth claims about events can never be proved, but some truth claims can be demolished through procedures of contradiction.

If data belongs to an event, it cannot be attributed to a particular agent or aspect of that event. It is common and comforting to attach data to particular objects or participants in an event, and to the extent that all other participants and relations that constitute the event are held constant and made overt, to that extent attributing the data to a particular agent constitutes a valuable shorthand in description and discourse. For example, to attribute a certain tensile strength to a steel beam is convenient, but has meaning only in regard to an event at which, at a certain temperature, the beam is stretched in a machine until it breaks. The time span within which this (hypothetical) event generates the same data is quite long. But over a thousand years, the steel beam no longer has this property; which is shorthand for saying it will behave differently in the event that it is stretched. Not only that, but any engagement in events will affect the tensile strength in an unpredictable way; if an unbroken part of the beam is stretched again it will be found to have a different tensile strength; as it will after multiple vibrations as part of a bridge.

So experiments in the physical and biological sciences do not produce data about the object, or measure properties of the object being investigated. They produce data about the event that is the experiment. Most experiments describe the behaviour of physical or biological objects under particular boundaried, that is, controlled circumstances. The information they give therefore is not so much about the "natural" world in which we and they live, as it is about the "controlled" world that is the experiment, and sometimes becomes habitualised as technology. Most social research has fallen into this trap of misrepresentation of the source and attribution of data.

Social events, or indeed interactional events of any sort involving living things, have time spans of small duration. Indeed, identical events are impossible to create because social relations, and the participants involved in them, continually change. Even if we could hold all the conditions constant as we do for the steel beam, the data still cannot be attached to the person because, even more so than for the steel, the person of tomorrow is a different person; and part of the difference is attributable to the experience involved in obtaining the data.

It follows from this epistemology that most psychological descriptions of people are shorthand and problematic descriptions of social events, from which most elements that constitute the event are camouflaged. The label is attached to the person even though the events which produced the data involved social interactions. This is an example of faulty labelling. In particular it applies to any notions of skill and competency that do not clearly define the context of their application.

So the issue of objectivity is not that things exist independently of the mind; the issue is whether things (elements) have properties independently of the events used to describe them. To say that a thing is real (has material existence) is very different to claiming that its "properties" are real and belong to it.

Ontology: What is the nature of social reality?

Within the meanings constructed above ontology precedes epistemology in that social relations are a particular case of an event in which two sentient beings (probably both human), are involved. By implication the event is the "reality." Something is happening "out there" that is producing a difference. Thus social experience is a particular form of experience of an event, and social meaning a particular construction of that experience.

On the other hand, epistemology precedes ontology in that all meanings are socially constructed, and are thus ultimately dependent on social relations and that includes the meanings we ascribe to ontology.

Regardless, the two domains interlink with no inconsistency in terms of the idea of social relations and the idea of knowledge being a function of experience of relational events, and meaning being socially constructed.

Using relations as a primary explanatory factor negates the notion of causality, at least in a simplistic sense. Events are construed as interactive systems where everything effects everything else; patterns of mutual influence replace causality as an explanatory principle. This has been generally accepted in Physics since the work of Einstein and Eddington early this century. It has always seemed odd to me that the more complex the system in which the event occurs - from physics through to biology through to social relations - the more frantically the idea of cause is clung to.

Further to that, the idea of "reality" is similar to the idea of "truth"; a redundancy, an unnecessary complexity, an irrelevant diversion. It contributes to conflict rather than to productivity. It seems more useful to talk about what aspects of social relations intrude most on experience, and are important to the intensity and duration of that experience, and the effects that it generates. In this regard I would make four assertions about social events, conclusions from my own experience and reflection:

- knowledge of social relations (that is, data generated within human interactions), is usefully construed in terms of the power and affect relations of the participants in the event; in particular, asymmetrical power relations generate different data than do symmetric power relations; and positive affect different data to negative affect (Foucault, 1988).
- an event occurs within specific localised power and affect contexts; this is not to suggest that this event might not itself be embedded in power relations (economically, racially, nationally or gender influenced) which push the effects and experience of the event in particular directions, but does put less emphasis on such grand power relations.
- events are dynamic, not static situations; they are characterised by movement, by change. They exist in time, which could be considered one measure of their change. So data about social interactions, which may often be characterised by power and affect relations, will change over time as the power and affect relations themselves change. I assume that any new social relationship (any social event characterised by people who have not met before in that configuration) will

initially be asymmetric in respect to power, and moot in respect to affect. The relational changes will affect the data generated through interaction, which includes discourse, and vice versa.

- Fixed societal structures (e.g., hierarchies) crystallise power relations and negate change. To the extent that they are successful they may produce knowledge, consensual interpretations, limited by the very boundary conditions that make its production possible; fixed societal structures also, in time, contradict the flow of interactional life, and produce social pathology.

Axiology: What values are embedded in the processes and product of the research? Whose interests are served through them?

No knowledge is value free. As Lincoln (1990) puts it, "given the criticism from all quarters, . . . only the most intransigent or the most naive scientist still clings to the idea that inquiry can, or should, be value free"(p82). Being socially constructed, knowledge produced from inquiry is related to the meanings and purposes and structures within which it was composed; and it will tend to confirm or negate those relations involved in its construction, depending on the interests and attitudes and assumptions and awareness of the researcher. Even if data could be produced that was independent of those elements and relations, that very independence is itself a value position, which could be construed either as objectivity, because it has transcended bias, or as ideology, because it camouflages the power relations from which its bias necessarily derives.

As a researcher my task is to contribute to the meaning system that helps me and other people make sense of their experience in the particular class of events with which this study is concerned. They will make sense of it if it is a story that links in some way with their experience, and at the same time is not contradictory to their experience; experience that is, of course, already partly interpreted in terms of other stories.

As an educator my task is to change people; education is nothing if it does not result in change. And as change is inevitable, but may be in many directions, there is obviously an obligation on the part of the educator to specify the direction in which change is intended.

As educator-researcher I must interact with the people with whom I wish to do research or educate. I do this through process (how I do the research), and product (what I produce as a result of the research). If I do not produce the data I investigate, but merely interact with data produced by someone else, this simply pushes the value problem one step backwards; their data was not value free. So if I accept their data without criticism, then I am accepting and perpetuating the values that affected its construction and effects. If I question that data, I question the social values embedded in it, as much as the social effects that are manifested through it.

If whatever I do involves interactions with people, and the construction of knowledge, then whatever I do affects both the meanings of people, and the social relations involved in those meanings. This is not to say that describing "what is" implies approval and acceptance of what is. Rather it is to claim that the very description of "what is" implies a way of viewing the world, a relationship with the situation, an involvement in the construction of the data, that pre-empts the meaning of the data by hiding the value assumptions behind the very mechanisms of its construction; becomes, that is, symbolic violence, unless made explicit (Bourdieu, 1977). Most quantitative research and much qualitative research is in this sense symbolically violent, in that the sources of its power are disguised.

Unless I wish to engage in a value contradiction, it seems necessary to have an awareness of the direction in which I wish to move people's overt and covert experience of social relations and the meaning systems construed within their influence; and to use processes and meanings that are congruent with those purposes.

My autobiographical note indicates that much of my work over the past thirty years has been involved with the nature and practice of violence in its various forms, especially as it affects young people.

My construction of the concept of structural violence (Wilson, 1992) indicates that I regard fixed hierarchical structures, in all their multifarious visible and disguised forms, as inevitably connected to structural violence and hence to social injustice. Due process within legal systems is necessary to alleviate, or control, some of the social fallout, but is not sufficient to ensure social justice at its root manifestation, which requires more equalitarian structures.

Peace and social justice are ideals that have many forms and faces that change over time. On the other hand, physical and structural and emotional and symbolic violence are constructs amenable to more specific definition, and hence more easily recognisable in particular social events. For this reason, I feel more comfortable having as a basic value the reduction of violence, which I could universally advocate, than with the increase of social justice, which is more nebulous because of its many-faceted nature; on this view, increase in social justice that is not associated with reduction in violence would be problematic, involving as it does an internal contradiction.

If beliefs (truths) are multiple, then so must be the values that are implied in those beliefs, or which inform them. How then can any particular value position be maintained as superior to any other?

In regard to the specific events that involve me and others in this thesis, I would answer that while the value of reducing violence is not necessarily superior to others, in the context of this work it is consistent with:

- 1. The learnings (culture and gender influenced as they are) that I have constructed out of my life experiences.
- 2. The ontology and epistemology which I have described, which inform the assumptions on which this study is based.
- 3. A view of life and living that involves ideas of growth, change, and flow at both individual and social levels. As such it is consistent with many views of personal enlightenment and social justice.
- 4. Processes likely to favour the survival of human life on the planet at a time when the technology is available, and primed to destroy it (Schnell, 1980).
- 5. That universal attunement and compassion which is one aspect of the experience described as mystical, as cosmic consciousness, or as the perennial philosophy, which transcends historical and cultural boundaries, and contains a sense of the sanctity of each individual person (Wilber, 1991).

Slotting into the social research field: How does this epistemology, ontology and axiology fit into the social research field as currently constituted?

Some doyens in the research game still regard qualitative social research as an exotic rather than a native plant, and as such something to be treated with caution because of its possible ecological effects on what had previously seemed to be a very secure and threat-free environment. Specifically, many testing experts still live in a positivist world (Shepard, 1991). As well, most teachers are quite convinced that their tests measure their student's attainments; the correspondence theory of knowledge may well be discredited, and philosophically empiricism may well have been dead for forty years (Smith, 1993), but in schools and colleges and universities and work places it is alive and kicking. However, a rich literature has developed from the debates involving qualitative research over the last ten years (Burgess, 1985; Eisner & Peshkin, 1990; Guba, 1990 Popkewitz, 1984; & Smyth,1994).

So with some reservations qualitative research is now accepted and respectable, even though practice severely lags theory. The reservations are currently crystallising as sets of questions and answers about how to recognise "good" qualitative research. For example Carr and Kemmis (1985) describe five formal requirements for any adequate and coherent educational science (p158). Criteria and caveats are being constructed that will undoubtedly in time result in a new orthodoxy (Lincoln, 1990). Feyerabend's (1988) assertion that "science is an essentially anarchic enterprise; theoretical anarchism is more humanitarian and more likely to encourage progress than its law-and-order alternatives"(p5), provides as much discomfort in the research world, be it quantitative or qualitative, as in the world of politics or the family. Smith's (1993) work clearly indicates that clarification of the problem of criteria is central to any real progress. It is also necessary if any substantial change in educational practice, and associated structural relations, is to occur.

At this point in time, however, the limits of the field are blurry, and the demarcations between various camps subject to border skirmishes. So at least one reason for my position not fitting into a specific ontological, epistemological, axiological, or methodological tent is that such tents are not clearly differentiated between the encampments. Having said that, it is possible to nominate some camps to which I do not belong, and some camps to which I partly belong, where I would not feel too uneasy sitting in some of their tents.

It is generally agreed that there are three basic positions; empiricist (post positivist), interpretivist (constructivist), and criticalist (Smith, 1994; Lincoln, 1990). It is also agreed that this is an over simplification.

Briefly, empiricists argue that there is a reality out there to be discovered, that it is single and measurable, and that causal laws explain and predict it (Smith, 1994).

Carr and Kemmis (1983) characterise the interpretive approach to social science as aiming "to uncover the meaning and significance of actions" (p92). The interpretive position is that truth is constructed by people, and always involves a social context and social interactions. So truth is relative and multiple. This position has two strands, the ethnographic (Sherman & Webb, 1988), and the ontological strand (Eisner, 1988). The difference is in the way hermeneutics is regarded. In the ethnographic strand, hermeneutics is a method of achieving interpretive explanation; in the ontological strand hermeneutics is more concerned with the idea that all knowledge, all representation is dependent on the primacy of experience (Schwandt, 1990). Regardless, "hermeneuticists of all measure and variety agree that any interpretation of meaning must take place within a context" (Smith, 1993, p16).

Carr & Kemmis (1983) regard post-positivist and interpretivist accounts to be similar in that "the researcher stands outside the research situation adopting a disinterested stance in which any explicit concern with critically evaluating and changing the educational realities being analysed is rejected"(p98). However, some constructivists (Lincoln, 1990), more recently advocate an abandonment of "the role of the dispassionate observer in favour of the role of the passionate participant" (p86). This is a position with which I concur. Smith (1993) elucidates other similarities and differences in the various positions:

> Interpretivists take antifoundationalism to mean various closely related things such as that there is no particular right or correct path to knowledge, no special method that automatically leads to intellectual progress, no instant rationality, and no certitude of knowledge claims. These are ideas, of course that interpretivists share at one level or another with postempiricists and critical theorists (p120).

He goes on to point out that "differences of consequences are readily apparent as these points are elaborated upon more specifically"(p120), and presents his own view that

> the demise of empiricism means that it is time to move beyond the need for a theory of knowledge and the various dichotomies . . . of subject versus object, facts versus values . . . this is in marked contrast to attempts by post empiricists and critical theorists to elaborate a successor theory of knowledge by either modifying or recasting, respectively, the empiricist understanding of these dichotomies (p120).

The criticalist position also has two strands. In the first belong critical social theorists, ranging from traditional Marxists uncovering the "contradictions of economic conditions and relationships", to a variety of other critical perspectives, where "the focus is on the ideological distortions inherent in a broad range of historically formed social and cultural conditions" (Marshall, 1990, p181). Smith (1990) sums up the critical theorists project: "critical inquiry can reveal our objective historical conditions: tie this knowledge to the expunging of false consciousness, distorted communication, and so on; and thereby promote emancipation and empowerment" (p193). Critical theorists then have a clear agenda of social transformation, based on a particular historical perspective, to which they have appropriated the "objective" label. As Carr and Kemmis (1983) express it, they aim to "reawaken the power of criticism and the power of praxis - criticism and praxis being the critically enlivened forms of what we usually refer to as theory and practice" (p186).

The other strand of the criticalist position is the post-structural, post-modern strand, which includes some feminist perspectives. The concentration here is on the construction of social reality through language and discourse, and the way in which this serves dominant groups and interests. The emphasis in research is on discourse analysis, in order to expose such inequities (Smith, 1994). Foucault's work is sometimes attached to this strand, though he himself did not accept the classification. And I would agree. This is important, because the writings of Foucault considerably influenced this study.

So where does my position fit into all this? I am not a positivist or empiricist. I do believe that empirical data can be collected about events; it's just that I don't believe that in relation to social events such data is very stable, can be replicated without considerable error becoming evident, or can be justifiably attached to a particular participant constituting the event. Any such data views that event from a particular position, with particular boundaries, with particular interests and values influencing the collector.

On the other hand truth claims are sometimes explicit, and often implicit, in theoretical formulations or interpretations involving social events. And some such claims can be directly contradicted by empirical data, by effects or consequences that are directly observable.

In terms of ontology, of the nature of reality, I do not fit neatly into any of the camps; empiricist, interpretivist or critical. I am probably closer to being a sceptical mystic. Rather than enter into that potential bog, in this thesis I have bypassed the question of "reality" and begun with the notion of social events, which involve the participants in social experiences.

I am constructivist or interpretivist in as much as I see all knowledge as multiple and constructed. Eisner (1990) agrees that experiences are the basis for cognition and knowledge: "thinking and knowing are mediated by any kind of experiential content the senses generate...our language refers to referents we are able to experience, recall or imagine"(p91). However, as Schwandt (1990) points out, this ontological basis of experience is not common to all interpretivist methodologies.

Perhaps my main point of departure from the criticalist perspective is at the ontological level; certainly I see relations as fundamental in as much as they constitute the mechanisms through which difference and change occur, thus making events experiencable. But I do not wish to "objectify" these into some grand historical schema on the one hand, nor overemphasise their dependence on gender relations or particular discourses on the other. Rather, I see power and affect relations as a "heuristic fiction" that has great generality and elegance as an explanatory and generating principle. However, I am clearly allied with them in their wish to reduce the violation of persons through the transformation of social structures and in seeing social research as a legitimate way to help people make sense of the social world in a way that gives them some leverage to change it for the better. By "better" I refer to a decrease in violence.

A model for the assessment process

This thesis is concerned with a particular type of social event called assessment. It is particularly concerned with the assessment of individual persons. I assume that such an assessment results in a categorisation of some kind. Such a categorisation involves a bifurcation of data, itself dependent on judgments about criteria and standards.

Given the ontological position of the above discussion, the assessment process involves (at least) five stages (events) and a context. In actual practice some of these stages may be omitted or fused. Such fusion or omissions may constitute a source of confusion or error.

- 1. Test production: An event (experiment, test) is devised to produce data. Such an event will involve an interaction between the assessed person, and instrumentation of some kind. The instrument may exist in the assessor's head, or may be produced as a physical artifact (a written test). The test production process also involves explication of a theory-practice link of some sort, and some prior judgments about a relevant task.
- 2. Test experiment: The person being assessed does the test, by performing what is required in the testing situation. This is the first stage of data production, and this event is completed when the test is completed.
- 3. Data production: The second stage of data construction occurs when the assessor interacts with the testing process directly, or with products from it. eg. a performance or a completed test paper. This interaction involves an interpretation of the data.
- 4. Judgment process: This results in a categorisation of some kind; it involves a comparison of the data with the standard, either directly, or by comparing with data about other students. This process assumes the existence of the standard as a stable and replicable element in the event.
- 5. Labelling process: At least two labels are involved; the name of what has been assessed (described), and the name that describes the level of performance (compared to the standard). The multiple label is constructed from the whole assessment process, and is legitimately attached to those events. In practice it is more likely to be attached to an element of the testing event (the assessed), or to an even more remote theoretical construction related to the assessed (some skill or ability).
- 6. All of these processes are embedded in relations of power which reproduce and invigorate themselves in the processes. And all of these processes (events) are potential sources of error and confusion in the individualised material product of this whole process - the documented labelling and categorisation of the assessed person.

Summing up

Negating notions of truth and reality does not necessarily lead to chaos or alienation, but may presage a search for greater clarity of assumption, for greater precision of value, and hence for greater wisdom in action.

Part 2: Context

Chapter 4: Power relations
Chapter 5: Power relations in educational systems
Chapter 6: Standards, myth and ideology

Chapter 4: Power Relations

Synopsis

Power is defined in terms of relational fields rather than of personal or role attributes, of power as ruler and ruled. Arendt and Foucault articulate the construct differently in that they differentiate violence from power. I choose a broad definition of violence as any violation of personhood; so both force and physical violence are subsumed as sub-categories of that construct; and violence becomes a necessary aspect of asymmetric power relations, inevitable in hierarchies.

The other side of power relations is now highlighted; the side that produces rather than denies, that constructs rather than destroys. That is, I deal in some depth with Foucault's (1992) assertion that "power produces; it produces reality; it produces domains of objects and rituals of truth. The individual and the knowledge that may be gained of him belongs to this production"(p194). In particular, I look in detail at what is produced through two specific mechanisms fabricated within asymmetric power relations: the processes of disciplinary power, regulated through surveillance and penalty; and normalisation, achieved through linear labelling and sustained through the cult of individualism.

I look briefly at some of the "scientific" disciplines, and the micro-cultures that sustained them and helped provide their assumptions, theories and data.

Finally in this section Bourdieu's construct of symbolic violence, and the notion of habitus through which it is humanly experienced, shows how difficult it is, when playing the game our culture dictates, to recognise its limitations.

Defining power

What characterises social life is affect and effect; affect refers to those aspects of relating that are characterised by polarities such as emotional closeness-distance, of like-dislike, of attraction-repulsion, of affiliation-separateness. These affect relations are apprehended viscerally, experienced directly through the body. In the vernacular, in the field of sense relations you "feel the vibes."

Power refers to those aspects of relating that translate influence, that make a difference, that have an effect. The actions of one affect the thoughts or actions of another. The poles of a power relation could be characterised by such descriptions as dominant-submissive, controlling - rebellious, have - want, strong - weak. So within the field of power relations, what one person does affects a second, which affects a third, and so on. Such effects ripple onwards and outwards from human interactions in patterns that are indeterminate; yet even so the patterns are sometimes decipherable and probablistically predictable, for the fields that affect the patterns are stable and translatable.

For example, in all cultures there are families, groups of people genetically related whose patterns of interaction are relatively stable, whose ways of behaving towards one another are consistently patterned; the parent influences the child, the parent's demands produce action, the power vector is from parent to child. Yet even so the child's behaviour must influence the parent's behaviour, if only to maintain the parent's controlling function. In this sense power relations involve mutual influence, even though normally asymmetric, and translated into action involve dynamic events.

Such events are acted out in power fields, such as family or school or workplace, where the rules of the game are understood, and the overall direction of action influence predictable. In this sense the influence is not so much person to person as role to role; the relationship of parent to child overrides the relation of the person Jack to the younger person Julie. For this to occur we must assume some mechanism for the learning of relational roles, for the internalisation of the power injunction. For if we locate the power in a relational vector out there in the space between, we must also explain by what psycho-social means people in the field are moved to act. More of this later.

Affect and power relations are not mutually exclusive; strong affect can generate high intensity in the field of power relations. And doubtless asymmetric power fields are capable of generating considerable affect, both positive and negative. Even so, the two notions are separate, the two fields initiate different experiential effects, and are associated with different states of consciousness. Love and power are not synonymous. And which is stronger is moot. Like Bourdieu (1990 a), "We leave it to others to decide whether the relations between power relations and sense relations are, in the last analysis, sense relations or power relations"(p15).

Regardless of their relative strengths, their confusion produces dysfunction in societal relations, and pathology in individual people; love that degenerates into power play destroys itself; and power that masquerades as love is a sickening violation. However, this is too large a contention to debate in this thesis, and is not directly related to our major theme (Laing, 1967).

To summarise, I have defined power relations as the dynamics of mutual influence. In most situations such relations are activated in fields whose pattern is perceived by those who enter the field in terms of role relationships, or less consciously simply as appropriate behaviour, a predisposition to act in a certain way. People engaged in such fields are both activated and constrained, but by no means wholly determined, by the role expectations or predispositions (habitus) which, for individuals at either pole of a power relation, are activated by their entry into the field.

So let's see how this definition fits into the historical meaning of such concepts as power, force, strength, and violence.

Power and Rule

Traditionally the essence of power has been rule and command; or alternatively the act of ruling and commanding has been attributed to a faculty called power. This need to dominate was seen as an instinct in man, a psychological necessity. Force and violence in social life was thus inevitable, for they were necessary components in the command strategies of a leader. Combine this psychological instinct with the social requirement that the first learning of civilisation is that of obedience, and the two poles of a largely unidirectional power relation are accounted for. To command and be obeyed is thus the essence of Power. And the basic building block for monarchy, hierarchy, and their complex transformations into the modern state has been constructed (Arendt, 1970, p36).

A look at any parliament in action, or a peep into any political party meeting, leaves little doubt that this paradigm of the fight for dominance is still central to the inner workings of government; certainly jostling for place in the political party pecking order is a major preoccupation of politicians, particularly of those who aspire to top positions. However, tradition also specifies an alternative power game.

This was the idea of representative government, where obedience is to laws that have the people's consent rather than to dominant men, and elected leaders remain dominant only with the support of the people. This second paradigm undoubtedly has a much wider gap between vision and practice than does the first, and a fundamental question of political science has always been about whether this is ideology rather than reality, a fairy story that disguises and soothes the experience of most people of powerlessness, of alienation. Regardless, in most modern states there is some balance, some checks within limits, of the power of the state and the tyranny of its accompanying bureaucracy, articulated through the opinion of the people.

Arendt (1970) argues that all government - tyrannical, monarchical, oligarchical, democratic, bureaucratic, or whatever, depends finally on the support, the "qualified" obedience, of the people:

> All political institutions are manifestations and materializations of power; they petrify and decay as soon as the living power of the people ceases to uphold them. . . (so) one of the most obvious distinctions between power and violence is that power always stands in need of numbers, whereas violence up to a point can manage without them because it relies on instruments (p41).

Arendt wants the word power to be reserved for the many, as distinct from strength, which is a property of the singular, a function of character or charisma or physical prowess. So an individual who appears to have power has it only in relayed form from the many whose support is needed. Whereas violence uses implements to multiply strength.

Power and structures

What characterises all of these notions of power is their attachment to particular agents, either singly or in groups. Power is a quality, a property, of an object or objects. But there is another way of viewing power:

> The major contribution of what one has to call the structuralist revolution consisted in applying to the social world a relational way of thinking, which is that of modern physics and mathematics, and which identifies the real not with substances but with relations (Bourdieu, 1990 b, p126).

Bourdieu postulates the existence in the social world of objective structures, in addition to symbolic systems, and independent of consciousness and desires of agents; structures which guide and constrain their practices and representations, which produce a predisposition to act in certain ways (p123).

Foucault (1988) also moves well beyond the notion of "Power - with a capital P - dominating and imposing its rationality upon the totality of the social body." In fact, Foucault goes on to say, "there are power relations. They are multiple; they have different forms, they can be in play in family relations, or within an institution, or an administration - or between a dominating and a dominated class" (p38).

Foucault (1988), like Bourdieu, uses the relational power structure as a fundamental explanatory principle: "The characteristic of power relations is that, as agents in the structure, some men can more or less determine other men's conduct, but never exhaustively"(p83). So power relations precipitate all "the strategies, the networks, the mechanisms, all those techniques by which a decision is accepted and by which that decision could not but be taken in the way it was"(p103). Or in retrospect, that's the way it seems.

Power and violence

Yet like Arendt, Foucault (1988) wants to remove coercion, brute force, from his notion of power relations. He says:

> A man who is chained up and beaten is subject to force being exerted over him. Not power. But if he can be induced to speak, when his ultimate recourse could have been to hold his tongue, preferring death, then he has been caused to behave in a certain way. His freedom has been subjected to power. He has been submitted to government. There is no power without potential refusal or revolt (p83).

Yet the man chained does have a choice; to scream or not to scream. And surely Foucault would himself argue that what is conceived as an "ultimate resource" is itself a social construction - more a production of the particularities of his cultural experience than of some "essence" of humanness. And if so the difference he postulates dissolves.

Foucault (1982b) insists that

> What defines a relationship of power is that it is a mode of action which does not act directly or immediately on others. Instead it acts upon their actions: an action upon an action, on existing actions or on those that may arise in the present or the future. A relationship of violence acts upon a body or upon things; it forces, it bends . . . A power relation (demands) . . . the one over whom power be exercised be thoroughly recognised and maintained to the very end as a person who acts: . . (so that) a whole field of responses, reactions, results, and possible inventions may open up (p220).

In an otherwise articulate and logical essay on The Subject and the Power written at the end of his long career, Foucault in this passage seems to get lost. Actions now act directly on indefinite actions in an indefinite future in utterly magical ways; if power acts on the body it doesn't act on an action; the person at the dominated end of the power relation has to be recognised. By whom? Most of this is contradictory to all those subtle and unconscious "strategies, networks and mechanisms" through which he says the effects of power structures are promulgated.

There is some romantic idealism involved in this refusal to see violence as a special case of power relations, in this wish to make it a separate category. As Arendt (1970) admits, "nothing . . . is more common than the combination of violence and power, nothing less frequent than to find them in their pure and therefore extreme form" (p46). So what, if anything, is gained by making of violence a separate class of event? Is it that to separate them is to separate the human body, which can be subjected to the ravages of violence, from the "human spirit", which relates to power and can remain inviolate? This is a separation deeply ingrained in Western culture, which denies the integrity of the human organism, and wishes to separate body from soul, and nature (which includes woman) from man.

Perhaps both Foucault and Arendt, appreciating the necessity of power relations for all social functioning, and wanting to emphasise its positive constructive side, want to remove from its definition that which utterly negates the possibility of a spirited response; want to leave open the possibility of a political response in asymmetric power structures that are aided by overwhelming instruments of violence.

In other words, they reject a notion of structuralism in which only surfaces of humans, their bodies and behaviours, are involved; they wish to include the spirit, the internal meanings, as part of the equation; and the confusion arises from their own lack of clarity about how to slot in the subjective element.

Regardless, if we refuse to reify violence, and see it as a process, an interaction in which a living being is violated, then it becomes impossible to separate power relations and physical violations in this way, and it is clear that violations of an instrumental kind are but one strategy in a whole armoury of mechanisms available in the field of power relations for violating people.

Violation of personhood

Brown (1973) encapsulates this view in his definition of violence:

> The basic definition of violence (is) violation of personhood . . . And since personhood means the totality of the individual, and never just the body or just the soul, we are reinforced in our notion that violation of personhood can take place even when no overt physical harm is being done. In the broadest terms then, an act that depersonalizes would then be an act of violence, since . . .it transforms a person into a thing (p1).

So abuse, beatings, injury, torture and killing, what we normally recognise as violence, are more obvious forms of violation, and perhaps it is the intention to harm and the personalization of the act that makes such actions so abhorrent; the killing of a child with a bayonet seems more heinous than the more objectifiable destruction of a city with bombs. There is a different focus. Yet in the sum total of human misery and violation such intentional physical violence is minuscule.

People certainly are violated when abused or beaten or injured; yet just as certainly are they violated when disregarded or denied, infringed upon or intimidated.

People are disregarded when they are denied the basic rights of food, shelter or care, or full human status in communities. The mechanics of this disregard may be articulated through many systems, based on economics, class, caste, colour, gender, ethnicity, age, religion, or whatever; or more often some combination of these.

Denial, not recognising their existence as fully human persons, is one of the cruellest ways of violating, especially when perpetrated on young children, with its ultimate internalization of the destructive self image "I don't exist."

At a more general level, any positivist stance that treats people as objects, that directly or indirectly ignores of depreciates the internal meanings people create of events, is a violation of their personhood. On this basis much of current political ideology, economics, sociology, psychology, psychiatry, medicine, and educational and management practice, must stand condemned.

People are infringed upon in many ways: police or media or sexual harassment, smoke pollution in public places; confinement in school classrooms. Emotional or symbolic infringement is more subtle: a mother withdrawing love for disobedience; a preacher selling eternal insurance through inclusion in a particular group.

Intimidation also takes many forms; at its most obvious it is the threat of physical pain, at its more subtle the threat of hell. Intimidation feeds on fear; its father is the sword, its mother the imagination. Civilisation enshrines it in Law.

For the more sophisticated, intimidation is predicated on shame and guilt. Shame is the internalization of society's adverse verdict on behaviour, self disgust generated by what others think. Guilt represents a deeper internalization, the adverse criticism of self by self. Of all forms of human violation, the inculcation of guilt is perhaps the most oppressive, for guilt is pervasive in its influence and insidious in its effects.

In addition, humans are growing organisms. Their normal state is development, not stasis. So humans are violated not only when their physical existence or their psyche is threatened, but also when their capacity for growth is stunted, when their potential for expansion is diminished (Wilson, 1991, p16).

So we approach a dilemma: power structures are cultural necessities, the essence of community life, and at this point in cultural history all cultures are predicated in one form or another on asymmetric power relations; and all of the violations described above are manifestations of asymmetric power structures. It follows that violence necessarily flows from human culture as currently experienced. And attempts to separate power from violence involve inherent contradictions.

Power and production

One issue here is not whether asymmetric power relations predispose violations. They do. An equally important issue is whether they also have a productive role to play in the human condition. And they do. Foucault's great contribution has been to spell this out. "The refusal, the prohibition, far from being essential forms of power, are only its limits, power in its frustrated or extreme forms. The relations of power are, above all, productive" (Foucault, 1988, p118).

This view does redress the balance and help us to see the other side of the coin. People are produced and reproduced through their immersion in power structures. So are cultures. And the human spirit sometimes soars above the violence. Even so, the violations are often not extreme forms; they are inherently, pervasively and insidiously embedded into the structure.

So we must ask, what does "productive" mean in this context? If knowledge and people are socially constructed, what constitute the productive, rather than destructive manifestations of power relations? From what frame of reference is the separation between intellectual or emotional production and destruction recognised? As a starting point, let's first look briefly at Foucault's views about the mechanisms of this production, and then at Bourdieu's ideas about the inevitability of symbolic violence within reproductive cultures.

Disciplinary power

Over the past three hundred years, power on this planet has assumed a new face. Foucault (1992) traces this transformation brilliantly in Discipline and Punish:

> Traditionally, power was what was seen, what was shown and what was manifested, and paradoxically, found the principle of its force in the movement by which it deployed that force. Those on whom it was exercised could remain in the shade; they received light only from that portion of power that was conceded to them, or from the reflection of it that for a moment they carried. Disciplinary power, on the other hand , is exercised through its invisibility; at the same time it imposes on those whom it subjects a principle of compulsory visibility . . . the examination is the technique by which power, instead of emitting the signs of its potency, instead of imposing its mark on the subjects, holds them in a mechanism of objectification (p187).

Foucault is using the term "examination" here in its widest context. The written test as we know it is a refined and intense form of that "hierarchical observation" and "normalizing judgment" that characterise all examinations, whether they be pedagogic, medical, legal, penal, supervisory, psychiatric or whatever.

How is this power transmitted? What is the mechanism of its distribution?

> The power in the hierarchized surveillance of the disciplines is not possessed as a thing, or transferred as a property; it functions like a piece of machinery. And although it is true that its pyramidal organization gives it a "head," it is the apparatus as a whole that produces "power," and distributes individuals in this permanent and continuous field. This enables the disciplinary power to be both absolutely indiscreet, because it is everywhere and always alert, since by its very principle it leaves no zone or shade and constantly supervises the very individuals who are entrusted with the task of supervising; and absolutely "discreet," for it functions permanently and largely in silence. Discipline makes possible the operation of a relational power that sustains itself by its own mechanism and which, for the spectacle of public events, substitutes the uninterrupted play of calculated gazes (p177).

The details of this disciplinary power seem trivial in their manifestation:

> The workshop, the school, the army were subject to a whole

> micropenalty of time (lateness, absences, interruptions of tasks), of activity (inattention, negligence, lack of zeal), of behaviour (impoliteness, disobedience), of speech (idle chatter, insolence), of the body ("incorrect" attitudes, irregular gestures, lack of cleanliness) of sexuality (impurity, indecency). At the same time, by way of punishment, a whole series of subtle procedures was used, from light physical punishment to minor deprivations and petty humiliations (p178).

Together these trivialities articulate a milieu, produce an enveloping social environment, so that the people who live in that space accept it as a way of life, as a natural way of being. And so we find that, in the field of education

> A relation of surveillance, defined and regulated, is inscribed at the heart of the practice of teaching, not as an additional or adjacent part, but as a mechanism that is inherent to it and which increases its efficiency (p176).

Praise and blame

Disciplinary power uses the twin instruments of observation and judgment, and the judgment is by necessity judgmental; is categorised by a satisfactory- unsatisfactory dichotomy. Such normalizing judgments are so pervasive as to override their specific instances. "Humanistic" teachers may protest that they punish the misbehaviour and not the person; this may be true of their intentions, but does not describe the effects. Again Foucault spells it out; the judgments not only diminish the aberrant behaviour; they also produce the person:

> Through this micro-economy of perpetual penalty operates a differentiation that is not one of acts, but of individuals themselves, of their nature, their potentialities, their level or their value. By assessing with precision, discipline judges individuals "in truth"; the penalty that it implements is integrated into the cycle of knowledge of individuals (p181).

This translation of act into essence, of misbehaviour into attitude, of error into ignorance, of absence into inability, is one of the political functions of Psychology. This transformation of event into label is an epistemological error, a misrepresentation of the functioning process, but is crucial to the construction of those "individuals" of whom Foucault speaks. For as he indicates so clearly, that individual first constructed in the eighteenth century, that educated individual being continuously recreated in "developed" twentieth century countries, is not characterised by passion, creativity and an independent mind. On the contrary, the individual is a person cleverly moulded by disciplinary power to be utterly reasonable (that is, to deny emotion), completely responsible (that is, to deny spontaneity and creativity), and to be loyal and dependable (that is, to deny independent thought and action).

Illich (1971) reached similar conclusions:

> Under the authoritative eye of the teacher, several orders of value

> collapse into one. The distinctions between morality, legality and personal worth are blurred and eventually eliminated. Each transgression is made to be felt as a multiple case. The offender is expected to feel that he has broken a rule, that he has behaved immorally, and that he has let himself down (p32).

Normalizing

This process of creating the conformist and at the same time supporting the cult of the individual, is what Foucault calls normalizing. It involves five distinct operations. "The perpetual penalty that traverses all points and supervises every instant in the disciplinary institutions compares, differentiates, hierarchizes, homogenizes, excludes. In short, it normalizes" (Foucault, 1992, p183).

So what a child (or adult) does is seen not in its own right, but in the light of what others do. Behaviour and product, and ultimately relations and being, are constructed and thus perceived and conceived in comparative terms. So I do not exist in relation to others, but in comparison to them; I become an object in the field of comparison, rather than a subject in the field of creative and responsive relation.

The thrust of this comparison is not identification, but differentiation; the comparison focuses not on the similarities, but on the differences. The effect then is not to produce belonging and cohesion, but rather alienation and separation. And this differentiation is not in terms of the infinite variety of human behaviour and persona, but within a simple hierarchical catagorization of better or worse. To achieve this it is necessary to collapse the variety, the complexity, into a few single dimensions of value. And because the individual performances are indeed always multi-dimensional, and idiosyncrasies always do become visible, it becomes logically necessary to attach the value to the person, and not to the performance. The notions of skill, ability, attitude, intelligence, competence, morality, are uni-dimensional, and thus can be categorised and hierarchized as more or less, because they meet the joint requirements of unity and invisibility, and incidentally, of fantasy. (This argument is developed more fully in the chapter on comparability.)

And so we become homogenised, perceiving ourselves, and thus being ourselves, in the times and places constructed for us along the one-dimensional spaces into which we are constrained. It is as though hundreds of cakes, all made of different quantities of different ingredients, have to be rated in a competition. It is noted that most of the cakes expand on cooking. So we create a single variable called sponginess as a major dimension of comparison. Now we can proceed. The cakes are all more or less spongy. Now comes the moral shift. Some, indeed, are seen to be too spongy or not spongy enough. And so there evolves a notion of value within limits, of quality defined by conformity, of a homogeneity to which all good cakes must aspire.

These processes of comparison, differentiation and hierarchization lead necessarily to notions of the normal, of the acceptable, to the limits within which life must be lived, and outside of which punishments naturally accrue. The pervasive threat and final punishment is exclusion.

These modes of living are learned in most family settings, but the school classroom is the great levelling field where it pervades the life of the group. It is this pervasive quality that so affects the way of seeing other people and oneself that any other way seems alien.

In the late 1970s I was involved in a project in secondary schools involving non-judgmental assessment of students. That is, assessments that simply stated what they had done without that statement containing overtones of satisfactory-unsatisfactory, good-bad.

We explained to over a hundred teachers what we wanted. We asked them to consider particular students whose work they knew well, and to describe some particular examples of their work in this way. We ended up with some two hundred descriptions, of which we hoped to use twenty in our report as examples of non-judgmental descriptions of student work. In fact, none of them was suitable. The teachers were simply unable to write such descriptions; they were unable to see their students (or their student's work) in other than normalizing terms.

Their reality, based on standards, nullified their best intentions.

Individualism

We must not confuse the individualism of our current society with that myth of wild west rugged individualism which is part of the American dream, and exemplifies the "Aussie battler," though doubtless ideologues might welcome the confusion. The individual differences we produce are characterised by creating levels within homogeneous orders, by categorising along linear dimensions of value, by dichotomising continuous performances.

The person's individuality is thus produced by placing him or her along a simple scale, good or bad, satisfactory or unsatisfactory, suitability or unsuitability along a number of dimensions. The individual becomes categorised, described, and indeed produced by the grade, the mark, and finally the profile, which becomes the true description of the shape of the person.

The disciplines

Before we look in more detail at how the formal examination fits into all this, and more specifically the part that the notion of standard has to play, it is useful to fit this development into an historical context. For life was not always this way:

> Historically, the process by which the bourgeoisie became in the

> course of the eighteenth century the politically dominant class was masked by the establishment of an explicit, coded and formally egalitarian juridical framework, made possible by the organization of a parliamentary, representative regime. But the development and generalization of disciplinary mechanisms constituted the other, dark side of these processes. The general juridical form that guaranteed a system of rights that were egalitarian in principle was supported by these tiny, everyday, physical mechanisms, by all these systems of micropower that are essentially non-egalitarian and asymmetric that we call the disciplines. And although, in a formal way, the representative regime makes it possible, directly or indirectly, with or without relays, for the will of all to form the fundamental authority of sovereignty, the disciplines provide, at the base, a guarantee of the submission of forces and bodies. The real, corporal disciplines constituted the foundation of the formal, juridical liberties. The contract may have been regarded as the ideal foundation of law and political power; . . . The "enlightenment," which discovered the liberties, also invented the disciplines (Foucault, 1992, p222).

Here then, brilliantly summarised, is the monstrous double bind that accompanied the introduction of parliamentary democracy, the genesis of that sense that all thinking people have of "with all these freedoms, how come I don't feel free?" And looking around, they do see all those economic, class, race, gender sources of inequality, and direct their attention to their amelioration, and forget that all were constructed out of the same structural cake mix, from the relations of disciplinary power embedded in hierarchy.

Yet there was a further development here that added immensely to the effects. The hospital, the school, and the workplace, once they had become located as gardens for the growth of disciplinary techniques, at the same time provided nourishment for the accumulation of new branches of knowledge. Clinical Medicine and Psychiatry became branches of knowledge predicated on hospitals and asylums; Education and Child Psychology were branches of knowledge predicated on schools; and Management Theory is predicated on offices and factories. (Offices are no less offices because their power relations and communications are crystallised through computers and their agents can be physically widely dispersed).

It is important to realise that these branches of knowledge developed after the structures, both physical and relational, were in place, and not the other way around. What we have here is knowledge developed within institutionalised relations; knowledge of people already objectified by disciplinary power; knowledge, that is, predicated on institutional inequity, and thus committed to rationalising that objectification.

So pedagogy is knowledge of the learning of children confined in classrooms, just as child developmental psychology is an accurate description of the growth patterns of children produced (both constructed and oppressed) in family and school. When the common translates into the normal and hence the real, these descriptive charactertures define the nature of children.

The unexamined givens of these systems of knowledge are the institutions in which they are based, just as the power relations that are embedded in these institutions comprise the assumptions on which these disciplines are built. And in its turn, the knowledge produces a magnification of that power asymmetry, both because it forms the basis of a verbalised truth that necessarily supports the institutional structure, and because it becomes the property of the professionals who practice it, thus necessarily excluding all others from its mysteries.

Ideologically, these disciplines claim to modify the negative effects of disciplinary power, which

> .seems to have undergone a speculative purification by integrating itself with such sciences as psychology and psychiatry. And, in effect, its appearance in the form of tests, interviews, interrogations and consultations is apparently in order to rectify the mechanisms of discipline: educational psychology is supposed to correct the rigours of the school, just as the medical or psychiatric interview is supposed to rectify the effects of the discipline of work. But we must not be misled; these techniques merely refer individuals from one disciplinary authority to another, and they reproduce, in a concentrated or formalized form, the schema of powerknowledge proper to each discipline . . .the examination . . .is still caught up in disciplinary technology (Foucault, 1992, p226).

Now perhaps we can begin to get a little glimpse at the forces that we are contending with here in the field of education. If Foucault is right, then the tenacity of the examination as an educational technique, no matter how professionally denigrated, is easier to understand. And if, as I shall try to show, the examination has no teeth, indeed becomes a paper tiger, without the notion of the standard to support it, then we begin to understand why the empirical facts about the instability, idiosyncrasy, non-transferability - in short, the factual non-existence - of the standard and its measure, has been so consistently and successfully suppressed and repressed.

In the following passage Foucault (1992) indicates the centrality of the idea of the standard. And whilst he is referring here more to standards of social behaviour, they apply equally to more cognitive matters:

> in the genealogy of modern society, they (the minute disciplines) have been, with the class domination which traverses it, the political counterpart of the juridical norms according to which power was redistributed. Hence, no doubt, the importance that has been given for so long to the small techniques of discipline, to those apparently insignificant tricks that have been invented, and even to those "sciences" that give it a respectable face; hence the fear of

> abandoning them if one cannot find any substitute; hence the affirmation that they are at the very foundation of society, and an element in its equilibrium, whereas they are a series of mechanisms for unbalancing power relations definitively and everywhere; hence the persistence in regarding them as the humble, but concrete form of every morality, whereas they are a set of physico-political techniques (p223).

Educators have been slow to appreciate the implications of Foucault's work to their own discipline. Foucault and Education (Ball, 1990) does explore this domain. And many of the contributors to this book identify the examination as the crucial stategy for embedding knowledge relations into power relations. For example, Hoskin (p31-32) and Jones (p84-97) identify the examination as the pivot of those small techniques through which the modern person is both constructed and controlled.

Symbolic Violence

Before discussing further the place that the examination plays in disciplinary power, I want to examine in more detail the notion of symbolic violence, and the particular way in which it is concerned in the continuance and intensification of violating structures through the imposition of meanings.

The child who is beaten by her father, and is then told that it is God's command that she must always love and respect her parents as indeed her parents love and respect her, and whatever they do is for her own good, is being subjected to symbolic, as well as physical violence. Her experience of being violated is being contradicted and negated. She is told that she is not being violated, but is being helped and loved. And it is not her parents who wish this, but God. She is unable to see that the perpetrators of the violence, and of the meaning system, are both primarily concerned to maintain their own, and each other's, authority structures; that is, the hierarchical power structures that have become institutionalised as family and church. And it is the institutions themselves, not parental love or god, that legitimise the violence, and the justification for it. So these structures become stronger, and the human victims more confused and powerless.

Let's take another example from schooling. Some young people are denied the right to continue their studies. Schools deny them access to further education and hence exclude them from a number of occupations. This is obviously a violation and unjust, even before we look at the inequalities of exclusion in terms of social class, gender and race. How is this exclusion achieved? Schools impose what specific knowledge and skills will be taught, and in so doing define what is useful and legitimate knowledge, and how it will be taught, learnt and assessed. And these processes discriminate against certain groups, and certain particular sorts of people.

The exclusions are legitimated supposedly through the professional judgment of the teacher, who is able to distinguish a "pass" from a "failure." In fact, this is not true. It is the institution itself, the school, that legitimises the exclusion, and inclusion. For the teacher outside the institution, no matter how highly qualified professionally, cannot accredit. On the other hand, the institution can accredit with a multiple-choice, computer-marked assessment system that completely bypasses the professional teacher. So what are in fact rather arbitrary impositions by the school are disguised as professional judgments about skill, ability, and intelligence, and then codified pass or fail with the appropriate label attached to the student. These judgments are then accepted as legitimate by all parties involved, including the great bulk of excluded students, who know at one level that they have been duped, but don't know how.

In these two examples I have tried to elucidate the particular properties of symbolically violent meanings. Firstly they are meanings imposed and legitimated by institutions of authority. For example, by institutions that control morals or education or health or information. Secondly they are designed to convince that what is violent is indeed not so. That what is unjust is indeed just. That what is inequitable is indeed fair. That is, meanings that are symbolically violent negate our experience and feelings. And thirdly, the authority appears to come from a source other than its true one. From God or some moral or professional source, rather than being delegated from less visible power structures of church, caste or class (Wilson, 1991, p26).

These are specific examples of Bourdieu's (1990a) more general proposition that

> Every power to exert symbolic violence, ie. every power which manages to impose meanings and to impose them as legitimate by concealing the power relations which are the basis of its force, adds its own specifically symbolic force to those power relations. . . . . All pedagogic action is, objectively, symbolic violence insofar as it is the imposition of a cultural arbitrary by an arbitrary power (p4,5).

Bourdieu shows that pedagogic action reproduces the dominant culture in two senses; firstly because the power structure within which the learning takes place tends to mirror and legitimate, and thus reproduce, that of the dominant culture; secondly because the meanings inculcated have been selected (with corresponding exclusions) to reproduce the meanings of dominant societal groups. Both structure and meanings are arbitrary insofar as the structure and functions of that culture cannot be deduced from any universal principle, not being linked by any sort of internal relation to "the nature of things" or any "human nature"(Bourdieu, 1990a, p8):

> The sociological theory of pedagogic action distinguishes between the arbitrariness of the imposition and the arbitrariness of the content imposed, only so as to bring out the sociological implications of the relationship between two logical fictions, namely a pure power relationship as the objective truth of the imposition and a totally arbitrary culture as the objective truth of the meanings imposed. (p9) . . . authority plays a part in all pedagogy, even when the most universal meanings (science or technology) are to be inculcated. There is no power relation, however mechanical or ruthless which does not additionally exert a symbolic effect (Bourdieu, 1990a, p10).

Habitus

When a person has "lived" long enough through a period of inculcation of training, there is a durable product internalised by them which Bourdieu calls a habitus. Durable because it remains after the training has ceased, and is capable of perpetuating in practice the principles learnt. In this way the habitus produces and reproduces "the intellectual and moral integration of the group or class on whose belief it is carried out "(Bourdieu, 1990a, p35).

The habitus is a system of schemes of thought, perception, appreciation and action, a predisposition to "a rule-bound activity which, without being the product to obedience to rules, obeys certain regularities" (Bourdieu, 1990a, p64). Bourdieu (1990b) uses the analogy of the game to explain how the habitus functions:

> The habitus as the feel for the game is the social game embodied and turned into a second nature. Nothing is simultaneously freer and more constrained that the action of the good player. He quite naturally materializes at just the place the ball is about to fall, as if the ball were in command of him - but by that very fact, he is in command of the ball. The habitus, as society written into the body, into the biological individual, enable the infinite number of acts of the game - written into the game as possibilities and objective demands - to be produced; the constraints and demands of the game, although they are not restricted to a code of rules, impose themselves on those people - and those people alone - who, because they have a feel for the game, a feel, that is, for the immanent necessity for the game, are prepared to perceive them and carry them out (p63).

So the rules of the game construct the players, who in turn construct their own particular version of the game. And those who play the game the best are the winners who continually reproduce the game in its infinite variety, and create the illusion of freedom whilst the rules become ever more fixed, for

> The pedagogic work which produces the habitus . . . produces misrecognition of the limitations implied by this system, so that the efficacy of the ethical and logical programming it produces is enhanced by misrecognition of the inherent limits of this programming . . . The agents produced by pedagogic work would not be so totally the prisoners of the limitations which the cultural arbitrary imposes on their thought and practice, were it not that, contained within these limits by the self-discipline and self-censorship (the more unconscious to the extent that their principles have been internalized) they live out their thought and practice in the illusion of freedom and universality (Bourdieu, 1990a, p40).

Bourdieu (1990a) here demonstrates how difficult is to question the principles of one's own culture, for the very questions have their roots in that culture (p37).

Summary - power relations and standards

In this chapter I have started to reveal the backdrop for our drama, those social and political fields in which the human actors are enmeshed. The focus was on power relations, and the way in which they both violate and produce those who act out their lives within their pervasive influence.

In particular the mechanism of disciplinary power relations was examined, and the part that the normalising gaze of the examination has in controlling the players, and creating the modern individual as its supreme production; an individual defined by a competitive profile, an object positioned, classified, and articulated along a limited set of linear dimensions.

In the next chapter I show that crucial to this extremely efficient mechanism for achieving social stability is the scalpel that defines the classification that produces the person that lives in the house that disciplinary power built. A scalpel labelled standard!

Chapter 5: Power relations in educational systems

Synopsis

In this chapter, I take the more general ideas about power relations discussed in Chapter 3 and apply them to educational systems and institutions; in particular I unearth the many small social control mechanisms that pervade the school, and what sorts of people are produced by those mechanisms. I then examine the examination; how it normalises and individualises, and how it is impotent without the notion of the standard, the sword that excludes and rewards, the wedge that produces the gaps.

That brings us to the focus of this thesis, the suppression of error. There is a field of educational scholarship devoted to educational evaluation and measurement. Thousands of books. Hundreds of Journals. Most of the literature in the field is about errors in measurement. And of course, errors in measurement imply errors in the measurement of standards. Yet in classrooms and universities and public examining boards, on school reports and graduation and proficiency certificates, there is a great silence. It is as though this literature did not exist. Even prestigious testing agencies skim the surface of the error issue. The question is why? Why this suppression of the obvious empirical fact that educational standards as a thin accurate line have no empirical existence? It is to this question that the remainder of the chapter is addressed.

I examine the crucial part that the standard plays in the whole mechanism of defining cut-offs for abnormality and non-acceptance, and how important it is that these standards be seen as accurate if current societal structures are to be maintained.

Restrictions, penalties, productions

In the day to day operation of the school the power relations are activated through an array of petty restrictions and micro penalties, unrelated to the supposed primary function of the school as an institution designed to maximise learning. In most classrooms the policing of these restrictions takes a considerable amount of teacher time and often consumes more physical and emotional energy than does their teaching function. In many large High Schools in Australia, the major activity of the Deputy Principal is to deal with children with whom teachers are having disciplinary problems. We are obviously dealing here with what is a major part of the school curriculum, regardless of whether it appears in the official statement of syllabus.

There are restrictions on appearance and dress; on what may be worn, and how long or short it is; whether this be skirt, shirt, pants, hair, necklace, ear rings - whatever differentiates from the norm; whatever distinguishes an idiosyncratic persona; whatever, by whatever means, makes a public statement about personal autonomy. The restrictions will not be specified in detail, for fashions change too fast for that, and student creativity is limitless. However, the judgment of the school is, in retrospect and by definition, impeccable in these matters, and their verdict will rarely be contradicted, and never successfully challenged, by students. (or parents, for that matter). Significantly, school spirit, cooperation, health and safety, economy, equality, fraternity, are all likely to be part of the supporting ideology. But never conformity, for this would contradict the school ideological aims of developing individuality and autonomy. Yet surely conformity is what is being produced here; conformity, and the acceptance of the social sanctions that non-conformity bring.

Body, movement, speech and relations must be decorous: body and clothes must be not only clean, but tidy. Movement is both restricted and restrained: students should remain seated and never run in the corridors. Speech should be proper: slow, well-articulated, free of slang, swearing and salacity, respectful in address and tone, and preferably in the dialect of the upper middle class. And social relations should be moderate, free of all excesses; of love or hate, of enthusiasm or alienation, of spontaneity or cliquishness, of autonomy or dependency.

As well as physical and emotional containment, there is temporal curtailment. Work is restricted to what the timetable dictates. Maths must not be done in the history lesson, history must begin at 10 am., and no one may visit the toilet until 12.50 pm, unless they shame themselves by asking permission, and then only maybe.

There are a whole range of penalties utilised to reassert the power structure should any of the multitudinous restrictions of the school be breached: further physical containment during recesses, deprivations of various sorts, petty humiliations such as standing in corridors or outside offices, threats and harassments of various kinds, and finally physical punishment, suspension or expulsion. In 1997 in Australia the most popular fashionable sanction is called "time out", a broad notion that contains various shades of physical isolation, and which schools insist is not a punishment. The penalties are really of no significance. It is the acceptance of the penalty, which reinstates the integrity of the power structure, that is important. It is important that some students rebel, so that the power relations might be demonstrated (Wilson, 1990).

So what is produced through these restrictions and penalties? What is learnt? First, temporal regularity. There is a time to start and a time to finish, a time to sit and a time to stand. And these times are planned and arranged and policed by others. What is learnt is that time is determined not by the imperatives of life as they manifest themselves, nor by any plan that might make for some personal production, but by the dictates of people in authority, by the demands of an institution.

Second, physical containment. There is a space to be and a space to sit, and sit, and sit. What is learnt is that the demands of the body are not important, and it is preferable to forget that you have one.

Third, emotional contraction. What is learnt is that the exuberant emotional and psychic field must be reduced to the physical limits of the body, so that feelings and emotions are pacified, and the self reduced to placidity.

And finally, what is learnt is that all this has nothing to do with the maintenance of power relations, or the production of a social being, but is an unfortunate addendum to another far more important purpose; a necessary prerequisite for effective learning of the knowledge specified in the school curriculum. What is learnt is to misrecognise the social function of schooling.

Illich (1971) summarises the situation, calls it for what it is, and sees only one solution:

> School prepares for the alienated institutionalization of life by teaching the need to be taught. Once this lesson is learned, people lose their incentive to grow in independence; they no longer find relatedness attractive, and close themselves off to the surprises which life offers when it is not predetermined by institutional definition. And school directly or indirectly employs a major part of the population. School either keeps people for life or makes sure that they will fit into some institution. . . De-schooling is, therefore, at the root of any movement for human liberation (p47).

The examination

Before accepting or rejecting Illich's ultimate solution, let's look more closely at some of the specific mechanisms that produce this "alienated institutionalization of life."

First we look more closely at the examination, and at the particulars of its function. Foucault (1992) certainly affords it pride of place among the mechanisms of disciplinary power which he elucidates:

> The examination combines the techniques of an observing hierarchy and those of a normalizing judgment. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualized. In it are combined the ceremony of power and the form of the experiment, the deployment of force and the establishment of truth. At the heart of the procedures of discipline, it manifests the subjection of those who are perceived as objects and the objectification of those whom are subjected. The superimposition of the power relations and knowledge relations assumes in the examination all its visible brilliance (p184).

The examination is the ceremony of ordering; it is the mechanism through which real people (and hence the world) is ordered, and held in order, in all of the meanings of that word. By doing this in a setting in which the person who establishes order is also the person who establishes truth through knowledge, the certainty of correctness is established, and the person becomes an object in the acceptance of their place in the line, in their acceptance of their uni-dimensionality, in their incorporation of their relative merit as an essential part of their beingness.

Of course the examination is also a crucial element in the construction of human cognition. It defines what are true and false facts, what is right and wrong thinking, and what are the acceptable limits of intuition and feeling. But we are more concerned here with social categorisation.

The report is the place where such individuality is made official; here is the permanent record, uncorrupted by any possibility of error, of one's place in the order of things; of a person's history, present, and future distilled into a single mark; of a sign that evokes possibilities and defines exclusions; in the world of higher education and the world of work, here is the official indicator of who you are, what you are.

Foucault (1992) indicates that this individualisation through comparison is intensified as power disperses and abnormality increases:

> as power becomes more anonymous and more functional, those on whom it is exercised tend to be more strongly individualized; it is exercised by surveillance rather than ceremonies, by observation rather than commemorative accounts, by comparative measures that have the 'norm' as reference rather than genealogies giving ancestors as points of reference; by 'gaps' rather than by deeds. In a system of discipline, the child is more individualized than the adult, the patient more than the healthy man, the madman and the delinquent more than the normal and non-delinquent (p193).

It is at these crucial points that define exclusion that any error becomes unacceptable. These are the points that define, not so much the norm, but the gaps that define abnormality, unacceptability, dangerous deviance. The normal is indeed defined by a broad grey band, but it is essential that the abnormal be determined by the thin red line that separates. And that line, that thin red line where the blood flows, is the standard.

Standards and swords

Foucault does clearly show how the battle lines are drawn up. He displays the deployment of troops and the strategy of the battle. With unerring accuracy he pinpoints the diversions and ambushes and the misinformation and propaganda that camouflage the major thrusts.

Even so, he pays almost no attention to the major weapon which ensures success, to the one notion without which the whole structure is unstable; he downplays the construction that turns a house of straw into a house of bricks, and allows that momentous separation between the good little three little pigs, and the big bad independent wolf. Could it be that his academic self wished to retain this last bastion of its own identity?

Regardless, without the steel edged standard to cut off the tail with a carving knife, and without the standard chippy chippy chopper on the big black block to lop off the heads that are too way out, disciplinary power is reduced to a shadow. The notion of the norm is dependent for its existence on the notion of the not-norm, on the notion of the abnormal. And the abnormal owes its existence to the act of separation.

Regardless of how disciplinary power is deployed, whether through the micro-penalties of day to day detail, or the graduation rituals of national examinations, or definitions of insanity, the thin line between the acceptable and unacceptable must be drawn. And it can only be drawn by evoking the idea of a standard, of an cut-off point that can be accurately determined and applied. All this regardless of whether we want to evoke democratic values, or scientific values, or aesthetic values, or other "expert" values in determining the standard, and then measuring it.

For without the notion of the standard there can be no classifications, no qualifications, no exclusions. There can be no norm, because there is no abnorm. There can be order, but without the standard there can be no disorder; Without the standard, we can still construct an order of merit, but cannot differentiate excellence, or determine exclusion; we can still individuate by placing on a line, but we cannot delineate winners because we cannot define losers. A race where everyone gets a prize is like a race where no one gets a prize; it loses its purpose as a race, and soon becomes a game that no one wants to play. Gilbert was right: "When everybody's somebody, then no one's anybody."

The blade must be sharp. There is no room for error. There is some aesthetic beauty, some notion of swift justice, black and violent as it might be, in a blade that cleanly and swiftly decapitates. Yet a mangled hatchet job will inevitably evoke horror. And so it is with any application of the standard. The acceptance of classifications and exclusions, both by those who apply them and those who are their recipients, are dependent on the precision and truth of the standard. Without these qualities the whole examination exercise becomes exposed as a political ploy to order and control, to reward and exclude, to hold in place vast structures of inequity. In short, it becomes exposed as a hatchet job.

A place to hide

If it is indeed true that the notion of standard is central to the maintenance of cultural identity as we live it, as central perhaps as was the notion of God to the cultural identity of life lived in the Twelfth century, then we must not be surprised that the notion is highly resistant to empirical contradiction. Nor should we be surprised that those who are aware of any such contradiction have some realisation of its traumatic nature, and of the necessity to keep it secret.

The human mind is remarkably efficient. Socially inclined as it is, it realises the only way to keep a secret is to hide it away. So the secret becomes a secret from one's own consciousness, locked away down there where angels fear to tread. The unconscious is nothing more than this; the space where we hide what we know from our conscious selves because the knowledge contains a truth that is too hot to handle, an awareness too destructive to life as we know it.

Would the social world we know really collapse if the notion of the standard had to go? Would we dissolve in chaos, or move gently onward to build a better world? Or would we simply find another subtly socially reconstructed lie to replace the one we'd lost?

Summing up

We have seen how central the notion of standard is to the maintenance of the social structures of power in which we are enmeshed, and to education's crucial social function of categorisation.

There are affect components involved here; the bearer of the standard is clothed in fancy emotional underwear, wears a colourful mythical costume, and carries a sceptre that denotes moral high ground. In the next chapter we examine some of these other dimensions of the assessment fairy tale.

Chapter 6: Standards, myth, and ideology

Preview

After a brief look at myths and rituals, and the special place they hold in our thinking - a place apart from critical thought, I assign the idea of the human standard as currently understood to this mythological sphere.

I look at the emotional intensity of discourse about the standard, its significance as an article of faith, a basic assumption, an ideological king-pin, and at who gains from the non-recognition of its problematic classification. Specifically, I show how the notion of a standard of behaviour in families helps to maintain the family structure; then I examine in some detail the mechanisms the school uses to maintain "emotional" standards by denying the reality of human feelings, and how this is related to the maintenance of control, of good order.

Flags

When the army begins to march, or the Governor returns to his residence, the event is heralded by the raising of the Standard. The flag is the symbol of their power. When we salute the flag, we do obeisance to that power, in which glory resides. And, when power is embedded in the relationships of human structures, we salute the standard, we pay homage to the strength of those structures, simply by our willingness to play our designated part within them; in short, by our subservience to structural dictates, and our acceptance of relational obligations.

This language is hard to live with, this description too intense for comfort. We need a softer cushion on which to fall, a more prophylactic myth to justify our allegiances and comfort our losses. As we shall see, we will find such justification in the world of moral values.

These relational structures often have no visual symbol to represent them, though particular versions of them proliferate in the form of corporation logos, school and family crests. These are usually of limited emotional impact. More successful have been brand names for clothes, where the image behind the symbol has been so successfully assimilated that not only are consumers willing to pay much more for the product, but are proud to become walking advertisements. Some Japanese corporations and some sports teams have managed to construct songs that fit the bill. But in general the "flag saluting" within families, schools and workplace has been accomplished more through particular discourses with words and body language than through responses to visual symbols.

Discourse and value myths

I use discourse here to describe not only "what can be said and thought, but also about who can speak, when, and with what authority. Discourses embody meaning and social relationships, they constitute both subjectivity and power relations"(Ball, 1990, p2). Discourses thus constrain the possibilities of thought, and are defined by what is absent from them as much as by what is produced through them.

So what are the key elements of discourse around standards? What are the words and phrases that trigger a "flag" like response? For whilst it is true that most social structures can, if necessary, muster some physical force - in the form of army, police, courts, psychiatric hospitals, masculine muscle - to deal with minor perpertations of the structure, the inherent strength of the structure is vastly greater than such disciplinary mechanisms that may be utilised. Just as in a crystal it is the individual molecular bonds which bind the crystal in its hard, rigid and determinable form, so it is the acceptance and actioning by each person of the appropriate relational roles between people that account for the maintenance and solidity of the social structure. So how constitute a symbolic reminder, a conditioning stimulus, a ritualistic nudge and wink, that stimulates and fortifies the memories of our proper relationships to those who lead us or are led by us, to those who love us or whom we should love, to those to whom dues are owed, or to whom we owe our dues?

The gross but honest dictates of parent-child relations are not effective with adults, or for most children for that matter, raising as they do so much overt rebellious reaction. "Do what you're bloody well told" does not trigger the appropriate response. The linguistic flag carries much more powerful symbols in its armoury. Looking upward, we see Duty, Loyalty, Respect, Discipline and Strong Leadership all emblazoned on the High Standard in gold letters. And looking downward, the cold sharp chisel of Efficiency nestles neatly in the caring hand of Institutional Love.

It is important to understand that once these abstractions are incorporated into a personal value system, so that they become part of a way of being, a way of institutional living, the ground of faith on which hierarchical life is premised, then dependence and obedience all become responses that inhabit moral high ground, for they are necessary to maintain, not the hierarchy, but the values in which it is now delicately clothed. And the violations they entail work efficiently underground in this hallowed space.

Further to this, the more intense and horrible the violations involved, the more pervasive and enduring the myths and values that provide the cover up and justify the carnage. The Freudian myth embodied in psychoanalysis regarding the sexual fantasies of children is a good example. The myth enabled child sexual abuse and incest to be disguised and trivialised for a hundred years, as we are only now beginning to realise; sexual abuse of the child became translated through therapeutic discourse to sexual fantasies of the child aimed at the adult (Masson, 1991; Miller, 1984). The myth of the glory of war has required the joint barrage of visual human slaughter on television, together with an appreciation of the probability of global nuclear extinction, to diminish its insidious hold on our thinking. And even now the monster will not lay down and die.

And there is another aspect of enduring myths that we must not forget. Such myths do truthfully represent a part of the human condition. Many children do sometimes act seductively towards their parents. There is a form of transcendence in the self sacrifice and comradeship that is a part of some men's experience of war. Yet when these myths are used to disguise the carnage, rape and pillage that are their major manifestations, then such myths become not the harbingers of truth, but their disguises.

What I am asserting in this thesis is that the myth of the human "standard" is just such a myth in the more "civilised" wars of structural violation in which our lives are embedded, wars no less destructive of human life and potential because their weapons are so insidious and subtle: Wars to which at this time in our history it is now appropriate to turn our attention, so that we may, in a non-violent way, bring about their cessation.

Standards and discipline

Talk about raising educational standards evokes intimations of glory and solidarity, of battles won and lost, of remembrance of our dependence on elite leaders and arcane specialists. Who talks of the shocking implications of lowered standards and the necessity to keep them high, and to whom do they talk? Who are the flag-bearers to defend us from the horrors of mediocrity, and the hellish consequences of the (inevitable) average? What do such utterances herald, and do what do they respond? (Wood, 1987, p214).

In the public arena, whether that be the political castle of public affairs, the media circus of public relations, the disciplinary field of the public service, or the common ground of the public house, talk of raising standards is invariably linked with the idea of better discipline. Contrarily, the cause of lowering standards is clearly tied in public discourse to soft leaders and the inevitable anarchy which that is fantasised to produce.

So "standards are also values to which people aspire or lament the decline in or lack thereof." (Norris, 1991, p335). People talk about raising standards when they perceive a slackness in the ropes of control, when they see a sloppiness infiltrating the verities of life, when they begin to be fearful about life's diminishing certainties. Talk of standards is talk about conservation, about protecting the past in its imagined superiority and security, and defending the future through strong leadership. "Discipline," "Respect," "Standards," "Leadership" are almost interchangeable words in a discourse that lauds the good old days and decries the soft underbellied freedom and license of the present. It is the language of the old talking about the young, of the powerful talking about the rest of the world, of the mind talking about the body, of men talking about women. And these days, let us be fair, of some women talking about men. By implication, it is discourse that defends appropriation and privilege, and the structures of inequity in which they flourish.

Suffering together

Heraldic and educational standards both also share a deep emotional component, digging deeply into the well of group identity that tribes and political parties, multinationals and nation states, know so well how to bring bubbling and boiling to the surface. We all know the clarion cries that activate the emotional unity that is evoked and manipulated by demagogues - the Fatherland, the Motherland, Our Land, Our Nation, Our Church, Our Family, Our Team, Our God, whatever its particular form. Words that recall our common heritage and our common destiny, and the myths and ideologies that surround that communality; we lose our individual and insignificant identity in the power and communion of the group, and are seduced into forgetting our fear even as we lose our freedom.

Through such languaging the notion of standards and their conservation becomes emotionally tied to our deep sense of wanting to belong, wanting to have our place in the social world. And of course, our place in the social world is dependent on the survival of that social world in which we have our place.

At the very least, discourse about standards will be emotionally charged. Talk of changing educational standards is like talk of changing the flag. It triggers all the fears of change in the social realities, be they ever so violating, for which the standard, and the flag, are symbols.

By insisting in this thesis that educational or ability standards have no empirical reality, I cut much more deeply into the social fabric. For such a claim not only undermines the standard, but also by association denigrates the social reality that it represents. The metaphor is not changing the flag, but destroying it, on the grounds that the social order that it pretends to represent is a delusion, very different to the one that it does indeed refer to. A delusion whose continuance, furthermore, is largely sustained through the emotional effects of the inviolability of its recurring symbol, the flag.

The person who destroys the flag is inviting extreme social response, for such is its emotional content that many people will identify this map with its territory. For them, to destroy the flag is to destroy the social order it represents, and thus to destroy their identity within that order. Emotionally, social symbol and social reality are contiguous. For many people, this contiguity overlaps and symbol and referent become identical. In this state of mind, cognitive arguments and empirical data have as much impact as falling animals crashing into rocks. As much impact on the rocks, that is.

In an analogous way, to criticise the notion of educational or job standards on the grounds that they cannot in practice be measured or logically sustained is to destabilise the symbol of the meritocritous society, the competitive capitalist order that it supports, and the cult of individualism that, almost alone, it defines and constructs. Emotionally, these four constructs - standard, competition, meritocracy, and individualism, are deeply intertwined. To threaten one of them is to threaten all. And to threaten all is to threaten each one of us, you and I and him and her. For it is to threaten that social order in which we all, in our own way, or more likely in a way that the structure has imposed on us, has found our place.

Fact or faith - the sociological imperative

So the standard is a social construct whose meaning is not dependent on any empirical evidence to support it. The flag is not a bit of cloth attached to a pole; it is an idea, a social construct, with which most of us, individually and in a group, interact in fairly well-defined ways. In a similar way, money is not a piece of paper with pictures and writing on it. It is again a social construct which most people are willing to agree has a certain meaning which includes an intense emotional component. But again, a social construct dependent on faith for its continuance. Lose that faith, and the value of the money evaporates.

Likewise the notion of a standard: It is a notion, an idea, a social construct that helps bind together the social structure that brings order to our lives. If, as I have suggested, it is a very fundamental construct, one which is central and crucial to other social constructs which in this time and place are thought to have particular value in constructing (and thus validating and justifying) the social relations in which our lives seem inextricably enmeshed, then even more reason for letting it alone, for not subjecting it to too critical inspection, for not undermining a fundamental article of faith.

Articles of faith do not need empirical evidence to support them, and are extremely resistant to empirical evidence that casts doubt on their logical consistency or their stability or their contradictions to other articles of faith. For articles of faith tend to develop around themselves other ideas and ways of relating that are reasonably consistent with them. These coordinations then constitute a way of living in the world, a set of habits that helps give a sense of stability and thus timelessness in a world in which change is inevitable on every street, and chaos is just around the corner. They constitute, in other words, what we call social reality. They might more accurately be called the social fantasies we construct and live that help make the conditions of our lives, and the lives of selected others, more bearable.

And if this cuddly teddy bear turns out to be a real dragon, destroying the lives of many more than it supports, then all the harder to slay it.

The psychological imperative

When we are dealing with the educational assessment of students we must add the teacher's psychological necessity for accuracy. At some level teachers all know how important their assessments are to the futures of their students. They all are aware of its use in social stratification, and its more negative function of the excluder, and the destroyer of personal dreams. And this mechanism operates through self exclusion as much as exclusion by any external force.

This is the load the assessor carries: for the students themselves usually accept the judgments made of them, and compose their lives accordingly. This is self imposed as much as it is dictated by any external agency. So through their assessments, teachers have monstrous effects on the future lives of their students. This is an acceptable load if the assessments are very accurate, and do in fact measure the capability of the student. But if they are enormously in error, what then? What is the psychological price of instigating massive inequity, enormous misplacement?

Instrumental value

The notion of "standard" has a particular function in the value conglomerate of respect-discipline-efficiency that is a major part of the ideological glue that helps hold hierarchical systems firm. For the standard is the value that mediates between ideology and structure, between the moral values, and the relational power systems that they support. The standard defines the point of action at which any disjunction between value and experience is challengeable.

Let's see this in action in two hierarchies; first in relation to respect in the home; then in relation to emotion in the school.

The family

In a family, duty, obedience, respect, discipline are continuous, rather than binary, constructs. That is, children are more or less dutiful, or obedient, or respectful. One child is more disciplined than another. So how do we know when we reach the point where acceptability is breached, where unacceptability is reached? We know because what has occurred is below the standard. As parents we "know" there are standards of behaviour that must be observed. And the disciplined child is one who knows, accepts, and behaves within the limits of these acceptable standards. And these standards are not of my making as a parent, but something that "society" demands. I may have very high standards, in which case I may be tougher (and hence more moral) than most others. Or I may be softer (and hence more humane or emotional) than most others. But the myth of a "standard", that point of demarcation between acceptable and unacceptable, is implicit in both these positions. And my duty, as a parent, is to maintain this standard.

That this standard has no empirical stability (certainly not for the group and generally not for the individual) is insignificant in the light of its logical necessity to maintain the structural stability of the family. After all, how can a parent ever demonstrate the extent of power difference if that difference is never confronted with an explicit, implicit, or fantasised challenge?

Sexuality and school

The hierarchy that is the school is much bigger and less personalised, so is harder to hold firm. So there are many standards of behaviour to hold emotion in check, and many standards of cognition with which to gain leverage on the mental processes. This is equally true for both teacher and student. We like to make an ideological separation between school discipline and the school disciplines, yet the processes by which each are engendered are similar if not identical.

So how are emotions in a school controlled through the imposition (or better still the personal incorporation) of standards? Firstly there is the professional standard of distance, of objectivity, of detachment. Emotional involvement, whether positive or negative, is taboo. Professionally the emotions are controlled by pretending that they do not exist. On the positive side the standard is that low level of affect described as "friendly interest." For young children this may be expanded to "fondness" unless you are male and the student is female. On the negative side the standard, the limit of negativity, is a low key sternness that accompanies correction. Essentially these low level affects are seen as acceptable nuances of cognitive behaviour.

Neither anger nor love have any place within the professional role of the teacher. To indulge either is seen as a breach of professional ethics. Such standards are justified by claiming that any relationship with students involving emotion would be dangerous to the students involved and unfair to the others. Dangerous because escalation could lead either to violent or sexual outcomes. An example of the catastrophic consequence justification. This disguises the stronger and more immediate danger, of course, which is to the stability of the power relations. Legitimate anger at the inequities hidden in that structure, or of love that transcends it, both pose fundamental threats to its continuance.

For the student in school emotions are also ignored. They have no place and so do not exist. Any acting out of emotions however is given high priority and the school disciplinary structures are immediately brought into play. The emotions are ignored, but the behaviour is punished. This is equally true regardless of whether positive or negative emotions have inspired the behaviour. Indeed, the school authority is much more comfortable with handling the acting out of negative feelings of fear or anger or revenge or envy than it is with any overt expressions of love or sharing or student cohesion, so easily interpreted as solidarity and hence politically suspect as potentially destabilising.

Emotional intimacy between students, or between a student and teacher, is rightly seen to be incompatible with the power relations that define the school structure. Two students who actively demonstrate their passion are likely to be dealt with more harshly (probably by expulsion) than are those who actively act out their hostility. Hostile students allow the school to demonstrate its own power. Loving students can only highlight the emotional vacuum of the school's structure; and incidentally expose the obsession with sexuality that underlies its prohibition. That the taboo is so seldom breached is evidence of the school's enormous power, especially so during adolescence, where for many students it is their major preoccupation.

Demonstrated or inadequately disguised love between a student and teacher, even if completely non-sexual in its overt manifestation, evokes a response amongst teachers almost as powerful as the response to incest. Outside the context of the school, love between people of different ages is an accepted norm, so long as the differential is not too great. Within the school context, it is condemned on the grounds that it is an abuse of power. The assumption is that the teacher has abused his or her power over the student and manipulated the student's affection. Now whilst this may be true in some circumstances, and whilst the roles in the school have doubtless influenced the relationship, intense emotional relationships that develop between the two people (rather than between their partial selves in role) are much more than this. They are as common and as intense and as potentially fulfilling as are such relations occurring in any other social context.

To understand the strength of the taboo we must understand that it is not so much the abuse of power that is involved here, but its elimination, its disintegration, its transcendence. Love and power are incompatible relations (Laing, 1967). Love is a state of openness and mutuality in which the other is accepted in his or her wholeness, where there is trust in the flow of positive affect, of cohesiveness. Control is the denial of such trust, and structures defined by hierarchical power relations are thus structures permeated by mistrust (Maturana, 1980). Hence the necessity to control and punish.

So love relations between a student and teacher are not taboo because they might lead to sexual relations, or because they are unfair to other students, or because they represent an abuse of teacher power, or even because they might represent a malicious manipulation of the teacher by the student. Or because of the many additional justifications for the taboo that we could construct and fantasise. All would possibly at times contain some grain of truth, and all would miss the target by rendering it invisible.

The fundamental immorality of such relations is that they are contradictory to the structure of the school, to its defining power relations, and are thus a fundamental threat to its continued existence.

It is equally important to understand that this fundamental reason for the taboo will be disguised in any particular case by evoking the concept of standards. The teacher is at fault because she has breached a professional standard of conduct which involves the abuse of power. The student will be at fault because he has not realised his vulnerability and has not allowed himself to be sufficiently protected by the benevolent authority which has defined the standards of student behaviour. Like so many rules in a school, this one, about loving teachers, does not appear in the rule book. Even so, no student would truthfully claim they did not know that it breached the standard of acceptable behaviour. And few would be able to rationally justify its abolition.

As described earlier, the appearance of the standard invokes an emotional response rather than a cognitive one. It bypasses notions of equity or justice that might grow out of a rational debate on the power-control issue, on the limitation of personal freedoms. It sidesteps any possibility of an ethical discourse by asserting that a standard has been breached, and thus by implication some act at the best unsatisfactory, and at the worst grossly immoral, has occurred. As the interpreter of standards, the school authority no longer seems to punish in order to defend its unequable structure. It now punishes in order to defend a high moral principle encased within "society's" standards. A violation of human rights has become a defence of all those things that "society" holds sacred, which become classified under the general rubric of "responsibility." And the use of the "standard" is the primary mechanism through which this mystifying ideological scam is accomplished.

Mind games

So far I have been concerned with discipline, with the way the school deals with unacceptable behaviour. Yet in educational discourse this is considered an unfortunate by product of the school's function. School discipline is defended not so much in its own right, but merely as a prerequisite to the maintenance of the disciplines. After all, the "real" reason children are at school is to gain knowledge, to become adepts of the various disciplines. Such learning, it is claimed, is dependent on the production of order, so that any control function that the school has is there to maintain the order that makes learning possible. Children are punished in school not so much for their own sake, though "god knows they must learn to be responsible for their actions", but rather for the protection of others. All must accept the discipline so that all may learn the disciplines.

Taken as an assertion about the nature of human learning, this is ridiculous. To assert that the best way for children to learn is to sit them down at desks in a teacher dominated classroom containing thirty or forty other children and change to a different topic every forty minutes is to deny most of what we know about the variety of learning styles and efficient learning environments. It denies a hundred years of research about how people learn.

Yet still the statements about good order, which in practice means being obedient and conforming, are central to the school philosophy. The reason is that such claims are not amenable to educational discourse. They are political statements, not educational ones. They are ideological statements designed to preserve the structure, and not therefore touched by empirical data. As articles of faith, as fundamental assumptions, they are flag waving slogans, amenable perhaps to emotional manipulation, but not to rational discourse.

All of which is not to deny that in an authoritarian-dependency structure, good order is necessary for effective "syllabus" learning to take place. It is, of course. But beyond that, and more pervasively, it is that structure itself that is inimical to learning. And it is largely in reaction to that structure that disorder occurs.

The ideology of order is necessary to protect those power relations from the dangers of rational debate, and the destabilising effect of empirical information that such debate might make visible.

Teacher stress

This ability of the system to protect itself from destabilising influences is nowhere better demonstrated than in the matter of teacher stress.

While teachers "stress out" in droves trying to maintain order, this is considered a second order phenomena. Their "real" function is to teach knowledge and skill, and school authorities consider it unfortunate that personal deficiencies on the part of the teacher might cause them stress.

In South Australia, "Stress Leave" is only available to teachers who are classified as "sick". Stress is a deficiency label attached to the teacher, a medical condition divorced from relational life. It may not be claimed by describing either the overt or covert violations within the structure of schooling, or by explaining it as attributable to professional or personal conflict with managers or students. The price of obtaining stress leave is the absolving of the institution for any part in its causation. (Section 30: (2A), Workers Rehabilitation and Compensation Act, 1986, South Australia)

Standards and destabilisation

We have seen how the notion of standard is a crucial ideological and mythical element in the hallowed structure of society. And an essential characteristic of the standard for that purpose is that it can be accurately defined and measured. In fact, standards can sometimes be defined and measured, but the errors contained in such measures are very large. I will show that they are in fact much larger than the massive literature on educational measurement and evaluation suggests.

Regardless, the notion of error is intrinsic and fundamental to any notion of measurement, and hence to any notion of measuring a standard as it is understood in the academic literature. Singer (1959) goes so far as to claim that "while experimental science accepts no witnesses to matters of fact save measurements and enumerations, yet it will pronounce no verdict on their testimony unless the witnesses disagree" (p101). So experimental science requires differences in measurements before it can decide what the "best" estimate of the measurement is, and the very notion of measurement is predicated on the notion of error. On the other hand any error in measurement is unacceptable if the notion of standard is to fulfil it's societal function in the categorisation of people. Who would accept failure or exclusion on the basis of a mark of 49 percent - plus or minus 15? Or even plus or minus one?

The simple professional and ethical solution is to attach an estimate of error to every application of a measurement of the standard, a habit deeply ingrained into practice in the physical sciences. However, this so contradictory to structural stability in the social world that to my knowledge the issue has never been seriously raised in professional debate about examinations, and when on rare occasions "ability" scores are presented as bands rather than lines they are based on reliability rather than validity considerations, so are gross under-representations of error; they are fudged instrumental errors, rather than errors in assessment.

Summing up

The standard is a crucial part of the assessment myth that is central to the stabilisation of power structures in modern societies. As such, attacks on its integrity, the naming of the gross errors attendant on its measurement, and explications of the violations to individuals that accompany its use, will be resisted.

Notions of standard have a very high emotional charge, and those who defend standards inhabit the high moral ground, as they defend the faith.

So challenges will be rare, and will be seen by most people as immoral, because they threaten the social fabric.

In the remainder of this thesis, one such challenge will be mounted.

Part 3: Tools of analysis

Chapter 7: Four frames of reference
Chapter 8: Equity, frames and hierarchy
Chapter 9: Instrumentation
Chapter 10: Comparability
Chapter 11: Rank orders and standards
Chapter 12: An inquiry into quality

Chapter 7: Four frames of reference

Synopsis

In this chapter four different frames of reference are defined; four different and largely incompatible sets of assumptions that underlie educational assessment processes as currently practised.

First is the Judges frame, recognised by its assumption of absolute truth, its hierarchical incorporation of infallibility; second is the General frame, embedded in the notion of error, and dedicated to the pursuit of the impossible, that holy grail of educational measurement, the true or universe score; third is the Specific frame, which assumes that all educational outcomes can be described in terms of specific overt behaviours with identifiable conditions of adequacy, and what can't be so described doesn't exist; fourth is the Responsive frame, in which the essential subjectivity of all assessment processes is recognised, as is their relatedness to context. Here assessment is a discourse dedicated to clarification, rather that the imposition of a judgment, or the affixation of a label.

Mythology

In the myth of meritocracy the examination is both a major ritual and a significant determinant of success. At the heart of this ritual, between the practice and the judgment, between the stress and the carthasis, is the great silence, the space where the judgment is processed.

The myth gives hints of what moves in this silence, for the myth makes three claims: the race is to the swiftest; the judgment is utterly accurate; and success is a certification of competency.

These hints tap the bases of the three frames of reference for assessment that assume objectivity. However, other assumptions of these frames make them mutually contradictory. This in itself would be good reason for keeping the process implicit. For the assumption that inside the black box hidden in the silence is a mechanism, an instrument of great precision, may be difficult to sustain, if it contains major contradictions within its workings.

Four assessment systems, with four different frames of reference, have staked their claim to exclusive use of the black box, their claim to be the best foundation for the precision instrument to measure human - what? Bit hard to say what exactly. To measure, perhaps, human anything. It may be sufficient just to measure. Or even just to pretend to measure, to assert that a measurement has been made, so that a mark may be assigned to a person.

Frames, myths, and current practice

The Judge's frame is far more often evoked than talked about. The focus is on the assessor's judgment of the product. The major activity is in the mind of the assessor. Such terms as expert and connoisseur are essential to the construction of the accompanying myth. Faith is the requirement of all participants. It is explicit in discourses about teacher tests, public examinations, and tertiary assessment, and implicit in all human activities that involve the categorisation of people by assessors.

The General frame is the basis for educational measurement, for psychometrics. The focus is on the test itself, its content and the measurement it makes. Such terms as reliability and ability are essential to its mythological credibility. It purports to be objective science, and hence independent of faith. As such the world it relates to is static, so there is no essential activity. It is explicit in discourses about educational measurement, standardised tests, grades, norms; it is implicit in most discourses about standards and their definitions.

The Specific frame is about the whole assessment event, and is the basis for the literature that derived from the notion of specific behavioural objectives. The focus is on the student behaviour described within controlled events; in these events the context, task, and criteria for adequate performance are unambiguously pre-determined. Reality is observable in the phenomenological world; the essential activity is what the student does. This frame is explicit in discourses about objectives and outcomes; it is implicit, though rarely empirically present, in discourses about criteria, performance, competence and absolute standards.

The Responsive frame focuses on the assessor's response to the assessment product. Unlike the other frames it makes no claims to objectivity; as such its mythical tone is ephemeral, its status low. This frame is explicit in discourses about formative assessment, teacher feedback, qualitative assessment; it is implicit though hidden in the discourses within other frames, recognised by absences in logic and stressful silences in reflexive thought. Within the confines of communal safety such discourses are alluded to, skirted around, or at times discussed; on rare occasions such discourses emerge triumphantly as ideologies within discourse communities.

The Judge

Most assessment in education is carried out within the Judge's frame of reference. The chief characteristic is that one person assesses the quality of another person's performance, and this assessment is final. By definition the Judge's assessment is free of error, and therefore any check of the Judge's accuracy would represent a contradiction of his function. So such a check is not only unnecessary, it is immoral, in that it is an act likely to destabilise the whole assessment structure by calling into question its most hallowed assumption.

The Judge's assessment may be verbal and on-site, eschewing numeration and a special testing context. However, performance is usually assessed with tests and examinations, with merit graded in some way. It is assumed that adequacy or excellence in performance is described accurately by the Judge. For this to be true, it must also be assumed that the test measures what it purports to measure, and that the marking, whether by the Judge or his assistants, is reliable. Again, therefore, checks of validity, that the test measures what it purports to measure, or of reliability, that the test will give the same result if repeated, are not only unnecessary, but are unacceptable and demeaning.

Judges must stand firm on the absoluteness and infallibility of their judgments, for this is the essence of their power, the linchpin of their role, the irreducible minimum of their function.

Thus they are duty bound to recognise standards, to perceive with unerring eye that thinnest of lines that separates the good from the bad, the guilty from the innocent, the excellent from the mediocre, the pass from the fail.

Talk to them of normative curves or rank orders or percentiles, all of which imply relative standards, and they will hear you out, wish you well, and with scarcely disguised distain send you on your way. In their absolute world such matters are irrelevant. They know what the standard is, and therefore their job is simple. Simply to allocate students, or their work, to various positions above or below that standard.

Set hard in a rationalist world view, this is a black and white world, a fundamentalist cognitive universe. The assumptions deny the possibility of reality checks, so the collective fantasy easily becomes the perceived truth, as human minds and bodies contort themselves to deny their more immediate experience.

So let us see what that more immediate experience might tell us if another frame of reference is chosen.

The General

The second frame of reference is called the General frame. I used to call it the generalizability frame, but that word has been hijacked by psychometricians. The general has been privatised and corporatised by mathematicians. The bird has been tamed and lost its wings. The general has become severely contained in mathematical armour.

What I am calling the General frame of reference is blatantly egalitarian and inherently relativistic in its conception, but has become constricting, reductionist and inequitable in its mathematical application. In one form or another it has dominated the academic literature in educational assessment for over sixty years. Within this frame is contained most of the received wisdom from thousands of studies in educational measurement and evaluation.

Its two initial assumptions are shattering. One Judge is as good as another. And all Judges are inaccurate. God is dead!

Now as Little Jack Horner understood quite well, you can't just stick in your thumb and leave it there. If you stick in a thumb you've got to pull out a plum or no one will say you're a good boy. And the plum was the third assumption: There is a stable rank order of merit. So there is a true score.

And there is a stable standard. It's just that, sorry old chap, it's just that the jury does it better that the judge. Or perhaps it would be more accurate to say that we measurement experts, we psychometricians, can do it, with the jury's help, much more accurately than you can.

Judge You can, can you?

General Yep.

Judge Whose assumptions are you using?

General Ours.

Judge Whose definition of a true score?

General Ours.

Judge Whose definition of error?

General Ours.

Judge And whose definition of standard?

General Ours.

Judge And you say I live in a fantasy world?

General That's what we say.

Judge I rest my case.

A bit unfair. But more that a grain of truth in all that. Even so, let's put a little more flesh on the skeleton of the General.

There is a true score: This notion has implications well beyond the psychometric. It is assumed that we are not measuring what a person can do, but rather a sample of what the person can do. If we could measure all the things (exactly) then we could find the true score directly. But as we can't there will always be some random error. In other words, if we had selected a different set of tasks the person would have done, probably, a little better or a little worse. Or even (softly now) a lot better or a lot worse.

This is all pretty obvious when you think about it. In almost any area of human activity, or study, there are an infinite number of possible tasks that could be required, questions that could be asked, limited only by the imagination of the examiners. And obviously, in a test situation, only a few may be chosen, from which a generalisation can be made about the rest. But the more tasks chosen, and the more they are a random sample of the total possible universe of questions, the closer you can get to the "true score". Further, *your* choice is a biased choice. Different people will choose different samples with different biases. So again, the more people involved in the setting of the examinable tasks, the closer we get to the replicable rank order, and hence to the true score.

We can't just stop at the questions, however; different markers rate answers differently. So markers also have to be sampled.

And contexts affect the result. Physical setting often affects performance. Some will perform better at home, some at school, some in an unknown environment. Some produce better work when isolated, as in a "normal" test situation. Others require stimulation in a group, which approximate more "normal" work situations.

The interactional media is sometimes crucial. Some express themselves better with the written word; others are much more comfortable with visual, aural-oral or more physical communication. Meanings can be communicated through many sensory modes. So if we are concerned to assess understanding of some area we would logically need to check across all of these modes.

And the time is important. They might do it well before lunch, badly after; successfully today, unsuccessfully in a month's time.

So assessments are required (marks or grades or rank orders), in all these different ways if we are to get a true estimate of a person's attainment or ability.

Whoops

Whadaya mean, whoops?

I saw that

Saw what?

Saw you pull that card out of your sleeve.

What card?

That one with the word "ability" on it.

I didn't pull it out of anywhere. I materialised it. I created it.

You made it up.

I created a useful concept. We all do it all the time.

Useful to who?

Useful to me.

Why is it useful to you to make up a concept called ability.

Because I've created a mess. A conglomerate of numbers based on myriads of interactional and contextual incidents. And I know how to turn it into one fairly stable number. But then I've got to write it on a label and pin it on someone.

Why?

Why?

Yes, why?

Well, if I can't pin it on someone then I would have done all that work for nothing, because it's obvious that although all these scores and grades were supposed to be measuring the same thing, they were actually measuring different things.

And you've got to have them measuring the same thing?

Obviously, otherwise I can't add up all the marks to get one stable mark, can I?

I suppose not.

So I made up a name.

Ability?

Ability.

And no doubt you specified the ability as being identical to the task area you were assessing?

Of course.

So ability is what the total (average) number is measuring?

Absolutely.

Relatively, you mean.

Yes, it would be fairer to say relatively.

And if you know their ability you know what particular things they can do?

No, I wouldn't say that.

Perhaps you know what particular things they can do better than someone else?

No, not that either.

What do you know then?

Well, if you were to take all the possible things that a person might be required to do in a particular area of activity that is more or less described by the ability, then you could say that, on average, and very consistently, a person with a high score on that ability would do better than a person with a low score.

Whoops, you've done another shift. All this information isn't about the person. It's about the interaction of the person with the task with the assessors. How are you justified in pinning it on the person doing the tasks? Why isn't this information about the whole contextual community?

Initially it is. But when we average out all the individual scores, they stabilise for each person. Regardless of the context, and regardless of the particular assessors. And the only other stable objects in the whole shebang are the people being tested, and the thing we're supposed to be measuring. So it makes sense. Ability is the stable label.

What does that ability score tell you about specific things that they can do?

In terms of specific tasks I would have to admit, if pressured to do so, that I could, from their ability score, predict very little.

So you began with lots of information about differences.

Indeed.

And you finished up with one bit of information and a name attached to a person. One bit of information about a constancy.

True.

You made a choice. You could have said that a student's true ability was all that variety of things that were very uneven and unstable and changeable. You could have said that the true description of ability was the collection, rather than the summary or summation, of all the information.

I could have done that.

And then the summary, the average, would represent a huge simplification, a reductionist symbol, a monstrous error, rather than a true score?

That follows.

But you chose to define the average, the summary, the abstraction, as the true score, and everything else as error?

Indeed I did.

How do you justify that?

Because the average gives a stable score, and a stable rank order, and this enables us to *make a clear classification of the student.*

And that's important?

It's crucial. You could say it was the aim of the whole exercise.

I thought the aim of the exercise was to describe a student's learning.

Would you think the best way to do that was with a number?

No.

Well, then!

I have tried to give some of the flavour of the General frame of reference here. To indicate some of its assumptions, some of the things it can do, and some of the things that it can't do. And it is apparent that one of the things that it can't do is give specific information about exactly what tasks a person can or cannot adequately perform.

I have also, in the spirit of this frame, fudged a bit. For example, the scores are not stable; they are stabler after they are averaged than they were before. As are the rank orders. But stabler does not mean stable; more reliable does not mean reliable; more valid does not mean valid. More of this later.

I have also expanded the conceptualisation of this frame well beyond most of the theoretical expositions in the literature. Such logical expansion does not lead itself to elegant mathematical modelling, however, so the fudging of psychometricians has reduced, restricted and simplified these concepts to a shadow of their full power.

The Specific

The third frame of reference for assessment defines the world of specific behavioural objectives, or specific learning outcomes, and, by implication if not practice, of the more fashionable criterion based assessment and competency standards.

Here we are far away from the religious world of the judge, and the pseudo-scientific world of generalised ability. Here is a technological space in which a spade is indeed a spade, and to Alice's delight, things are indeed what they say they are. Or so it would seem.

This frame of reference assumes that the task of assessment is to describe what can be done, under what conditions, and what constitutes adequacy. So there is only one correct description of performance, and that is the unambiguous learning outcome that is defined in advance. It is assumed that learning outcomes can be defined so clearly that there is no doubt whether a person has, or had not, matched behaviour to the outcome.

There is no problem here of matching objectives to curriculum, and curriculum to testing. The objectives are the curriculum are the learning outcomes are the test. A rose is a rose is a rose.

Here is the bright fluorescent material world of the technological fix. Reality defined as observable behaviour. A world where doubt and uncertainty is no more. A place of clear goals, purposeful activity, and attainable and unambiguous outcomes.

More than this. This is surely a political revolution. The power to certify or exclude is no longer in the hands of the omnipotent judge or the manipulative psychometrician. It is clearly with the student who can self-certify adequacy, and any intelligent bystander can check that the task has indeed been adequately accomplished.

The technique was first developed to train technicians quickly and efficiently during the second world war to do a limited number of very specific tasks, and follow through a finite number of carefully specified procedures. In this it was highly successful, and its overflow into the general training area, and the nebulous and vague syllabuses of education, was viewed with delight by many of those who wished a firmer base for guiding and assessing learning. That is, who wanted to control what people learn.

And it was possible to find in most areas of learning, in most specifications of jobs, in most definitions of curriculum, in most topics of study, some irreducible minimum, some particular aspects of performance such that we could say - well, if they cannot do at least these things to this level of skill, or if they do not know at least these particular facts, then we could never certify that they were adequate in this area of functioning. In other words, the frame proved to be very useful where there were a finite number of tasks that could be isolated and specified, with limits of adequacy defined.

However, there were two questions, one technical and one political, which shattered the image of specific behavioural objectives as a democratic panacea for education. The first question was - is it possible to specifically define outcomes in any area of interaction that includes cognitive or interactional areas involving any problem solving or analysis or synthesis. Any activity, that is, involving cognition of more complexity than low-level comprehension?

Note, however, that to ask this question is to step outside the frame. For the assumption of the frame is that all tasks are so specifiable.

And the political question - who defines the objectives? Why these particular tasks? Why this particular context? Of what significance this particular cut-off for adequacy? Have we solved the problems of reliability or adequacy, or merely hidden them behind a dense materialist behavioural smoke-screen, behind which shadowy judges, bureaucratically insidious, silently sit?

Again, to ask this question is to move outside this frame. Within the frame this question is not a contradiction, it is simply irrelevant.

The Responsive

The Responsive frame of reference for assessment is manifestly and covertly subjective: no longer are the descriptions and judgments attributed to the performance, the artefact, or the person. What the assessor says is no longer claimed to be a quality of the object produced, or the objectified subject that produced it. What the assessor says is claimed only to be what it indeed is - a response of the assessor to a particular situation or artefact; a verbalisation of a particular human response to an interaction; a construction of the person assessing that says certainly as much about the world view of the person assessing as it does about some abstract quality or behavioural skill of the object or person being assessed.

Within such a frame there is no question of a right judgment, of a correct classification, of a true score. The response might be sensitive or insensitive, sophisticated or ingenuous, informed or uninformed. The verbalisation of that response might be honest or manipulative, its fullness expressed or repressed, its clarity widened or obscured. It still belongs undeniably to the assessor, and the expectation is not towards a conformity of judgment, but a diversity of reaction. The lowest common factor of agreement is replaced by the highest common multiple of difference. The subject of assessment is no longer reduced to an object by the limiting reductionism of a single number, but is expanded by the hopefully helpful feedback of diverse and stimulating and expansive response.

As with the other frames of reference, this one rarely materialises in its pure form. In the evaluation literature it has gained some attention under the rubric of formative evaluation, which occurs during a course of study, a low status cousin of summative evaluation, the final judgment, that more macho space where the real battles are fought, and the important decisions are made. Even so, there is professional literature in plenty, and especially in the rhetoric of "teaching" rather than "assessment", that supports the idea of assessment as feedback and guide, rather than classification and judgment (Williams, 1967).

So it is in this diagnostic and formative function that responsive assessment has found its place; as part of the training program rather than as legitimate description of what has been learnt.

There is good logical reason for this. It is obvious that this frame is a direct contradiction to the Specific frame, in which there is only one description of performance required and that is defined in advance.

It is less obvious, but none the less true, that the frame contains, in its practical functioning, a contradiction of the Judge and General frames, for it denies implicitly the idea of the single accurate order of merit, and hence the notion of some true score, or of some inviolate standard.

There is a further contradiction built into the assumptions of the Responsive frame. For if, in attending to the feedback, the performance of the person assessed is indeed improved, then the quality of performance, the degree of skill, will be changed, and the "true score" will also be changed in the very functioning of the assessment process, making the accurate judgment immediately inaccurate.

It is important to the logic of the Judge, General and Specific frames that no learning takes place after the test, for otherwise the test result becomes invalid, and must surely be dispensed with. On the other hand, within the Responsive frame, it is expected that the responsive feedback from an assessor will interact with the performance and improve the quality of later work, at least in terms of that particular assessor.

In the Responsive frame, this is an act to be applauded; in the other frames, it is a worrying source of error; in this respect the Responsive frame fits into a dynamic, and hence educative, environment. The other frames are predicated on a static universe, and are thus, in a profound sense, anti-educational.

Shifting sands

How does the Judge perceive the other frames? To the Judge the General frame is hopelessly relativistic, lacking in authenticity and depth, and devoid of standards. the Specific frame is reductionist and trivial, unable to cope with the cognitive complexity which lies at the heart of any discipline. And the Responsive frame is permeated with that subjectivity that indicates the absence of the objectivity that only comes with true scholarship, which the Judge exemplifies.

How are the other frames viewed from the General perspective? The Judge simply cannot deliver his promise of measuring accurate standards. His idiosyncrasy is legion and his omnipotence is self delusion. The Specific frame presents information that is scattered, incapable of producing a single dimension of measurement. Any addition of the specific information loses it, and returns the data to the General frame without the usual measurement controls. The Responsive frame presents data that is too diverse and contradictory to be seriously considered as a measurement.

From the Specific frame the Judge may be measuring something but neither he nor anyone else knows what it is. Just so with the General frame, that gets lost in a wilderness of numbers and cognitive abstractions. And the Responsive frame belongs to the world of opinion and gossip rather than scientific description.

The Responsive assessor sees the Judge as a responsive assessor, deluded by a fantasy of objectivity and accuracy. The General frame is seen as mathematical chicanery used to justify unsustainable classifications of individual people. And the Specific frame is seen as an absurd attempt to reduce human experience and performance to a few describable and measurable behaviours.

Conclusion

Sensible debate within a particular frame of reference for assessment sometimes occurs. However, rational debate across the full range of frames is a rarity. Part of the reason for this is that people argue from different frames of reference, with their incompatible assumptions, and these are rarely made overt. Not only that, but individual people in a particular discussion shift from one frame of reference to another, sometimes with bewildering speed.

This is why a conversation between a university professor (Judge), a psychometrician (General), a educational software technologist (Specific), and a radical teacher (Responsive), sounds like the sound track from a Marx Brothers movie.

In the next chapter we shall see how these frames are related to concepts of equity and hierarchy.

Chapter 8: Equity, frames and hierarchy

Synopsis

In this section I want to tease out some of the relationships between equity and assessment.

Life wasn't meant to be easy. We have four frames for assessment. Four differing sets of assumptions about what assessment is about. Equity is similarly compounded. There are (at least) three differing definitions of equity in current use: The first is based on equal means, treating everyone the same; the second is based on equal ends, treating everybody differently to end up the same; and the third is based on elucidating different ends and different means. The advantages, limitations, and pre-conditions for these three notions to be effective in practice are discussed.

Then I take each frame of reference for assessment in turn, and tease out its compatibility with each notion of equity, and with the hierarchical power relations of which the assessment system is an integral part.

The meaning of equity

Equity means fair, says my dictionary. And fair means, you guessed it, equity. I asked my seven year old daughter what fair means. Sharing things, she said. Still not satisfied, I asked my five year old. Fair means not missing out, she said, being included.

That seemed like a good start. Notions of equal shares and inclusion. But the meaning gets more complicated as the implications for achieving fairness are developed.

Equal treatment

The soft definition of fairness is that everyone gets treated the same. But then they end up differently because different people respond differently to the same input. We can say that's fair because some people are more intelligent or work harder so we would expect them to gain more. But then if the nature of the input is changed, different people succeed. And the people who succeed often seem very similar to the people who design and implement the input. Not surprising really.

What has been designed here is a nice tight closed logical system; people design educational means and ends to produce people rather like themselves and also produce definitions of intelligence or ability or skill or relevant knowledge based on similar means and ends, thus justifying the fairness of the unequal ends in terms of the unequal intelligences of the people attaining them.

The self fulfilling prophecy continues when we make these unequal ends the criteria for selection to favoured occupations (Goslin, 1963 p156). Here the success of the incumbents, and all are deemed successful by definition once selected, proves the value and validity of the whole process. Certainly none of the people so favoured are likely to suggest that almost anyone could do their job given an appropriate training programme, or, even more unthinkable these days, through an informal apprenticeship.

How do teachers react to this soft definition of equity? For those who see their task primarily as transmitting certain knowledge and skill and attitudes to students the definition is appealing. Because they see their professional task as transmission, they are likely to define clarity of communication in terms of logic and intention rather than in terms of accessibility or effect. Thus their professional integrity will be preserved if they treat all students in exactly the same way. It will even be considered an advantage if all students dress the same in some sort of uniform so that personal idiosyncrasy is visually nullified.

At the other end of this spectrum are teachers, often those who teach very young children, who have some sense of the student as a person with a very particular background and learning style, and who have a sense of responsibility to deal with those differences, albeit with certain specified skills or knowledge as having particular importance. Such teachers will see the gross limitations of this equal treatment definition, and will tend to reject it.

Yet even these teachers are likely to be ambivalent about rejecting this definition entirely, because of their position in the total educational structure. After all, there is a curriculum that all students are expected to master, and the larger and more structured the organisational unit in which they are enmeshed, the more likely they are to feel the pressure and surveillance directed towards particular ends. And the bigger the group of students they are confronted with, the more helpless they are likely to feel about the possibility of treating everyone differently.

Then, confronted with the impossibility of treating the children differently, in confusion they abdicate: if it isn't possible to achieve equity of ends through differential treatment, isn't it best to at least achieve equity of means?

Equal ends

Let's take a closer look at this harder definition of fairness; fairness is treating everyone differently so they end up the same.

The reasoning is clear. People have different prior experience, so they necessarily start a new experience with different prior knowledge and skill. So if they are all treated the same, this differential starting point will produce disparate ends. It follows we must treat all of them differently if we are to give them all the same opportunity to reach the same specified end point. Fairness or equal opportunity thus means giving additional resources and time to those who are originally disadvantaged in order to achieve equality of ends.

Surely that's fair? Possibly. But who decides what these ends are that everyone should strive to reach? Usually they are defined by an unrepresentative group, who have a strong vested interest in maintaining and distributing certain sorts of knowledge, values, skills and myths, and/or of limiting the number of people who will have access to the same. Thus the ends are a narrow selection from a much wider range of possibilities. Why should all the resources go into these particular ends?

Part of the answer relates to the current nature of institutions, and the learning that can occur in them. They are not constructed or resourced in a way conducive to individualised learning, but in terms of much larger learning units.

So teaching institutions tend to ignore the unfair treatment of individual students for two reasons: First, because individual students have no power, this representation of unfairness is rarely articulated; and second, because an adequate differentiated response would administratively smell of disorder, such an approach would be contrary to the institution's structural purpose as a hierarchy, which is to impose order.

Some sub-groups however do have power. Institutions have to respond to claims of discrimination against particular sub-groups of gender, class, ethnicity, or whatever minority has found a voice. This has been useful in the short term as an awareness raising activity about the equity issue.

Such political activity on the part of sub-groups that have found themselves disadvantaged by current structures of teaching has resulted in some shift, at least in terms of rhetoric, towards the equal ends definition of equity. There has been some small acceptance of the idea that it is equity of ends rather than of means that should define equity.

However, the "equal ends" comparison has been applied to groups, not to individuals; the debate has been about whether as many girls as boys can join the power elite, and not about the individualised treatment that might allow all who so desire to be successful. As the debate is about the sharing of domination between groups, it largely ignores the domination within such groups. As such it is also about the sharing of violation, and not about its elimination.

Equal ends and the myth of the intelligent child

Action has been at two levels. One involves awareness raising, so that members of disadvantaged sub-groups are encouraged to attempt educational activities previously not sought; for example, girls to study mathematics or engineering.

The other action has been, not surprisingly, to attempt an economic fix. Just as economic health, on the current fashionable models, supposedly bears a long term relationship to standard of living and quality of life for all, so more resources for the "disadvantaged" sub-groups will supposedly produce more equitable ends educationally.

Such an approach ignores the relationship between means and ends. For if it is the means, in this case the particular form of educational environment, that has actually produced the different ends, then more of the same means is hardly likely to improve matters. Indeed, intensifying the same means may produce more discrimination. (Of one thing though we may be sure. More resources for the disadvantaged will certainly benefit those advantaged who have identified the problem, and have some solutions, preferably packaged.)

How could this be? How could an educational environment, created by professional teachers, produce negative results, increase disadvantage? Surely anyone with sufficient motivation and intelligence can succeed?

That's one myth that has always stood in the way of any real progress towards sharing and inclusion. Once you accept the idea of "bright" students and "dumb" students, and the notion that there is a direct causal relation between attitude and success, then inequities are merely a mirror of these individual variables. If girls don't do as well as boys it is either because they're not so bright, they're not motivated, or both. And poor kids are dumber than rich kids and that's why they don't do so well. It's obvious. It's genetic as much as anything. Rich kid's fathers are more intelligent otherwise they wouldn't be rich!

Teachers, armed with prejudicial expectations and judgments as well as assessment data, are often quite clear about who is bright, average, and not so bright in their class, a distinction not always so clear to the outside observer. I've talked to small groups of children in hundreds of schools. I'd often ask the Principal to select a small group of about twelve students, some bright, some slow (one of the in-words for stupid at the time). We'd sit in a circle on the floor in the library and talk about home and school and life and the future for an hour or so. At the end of that time I was never able to tell which of the students were supposed to be the "slow" ones. I suspected sometimes they included those who had made the most significant contributions, and the most profound comments.

The "blame the victim" ideology is pervasive in education, and is maintained through the closed logical system described earlier. Assessment procedures play a crucial role here. After all, the teacher is paid to teach. Yet the failure label is invariably attached to the student.

Different people, ends and means

Because both the common ends, and the means of attaining them, seem to contain within themselves the seeds of the inequalities we are trying to diminish, we can try a third definition of fairness.

Fairness is treating people differently so they can end up differently. And the different ends will be determined largely by the students themselves. Fairness than consists in providing different resources so that different people can achieve their own different end points, through their own appropriate means.

Is this individual choice and freedom not illusory? Surely expectations embedded in people's social class or gender will determine their choices, and so inequities of power and wealth will still be perpetuated?

This is not a light criticism, and the strength of such sub-cultural or individual expectations is great. However, this strength is diminished as the awareness and verbalisation of the imposed expectations increases. Sub-cultural expectations do not invalidate the logic of the "difference" definition. They do indicate some of the conditions for an implementation in accord with its purposes.

The professional rhetoric of education is concerned with ideas of "individual differences", of the "whole person," and of "clear thinking, rational man." Less so with the passionate, spontaneous, loving, emotional man, or woman. Even so, we might expect some professional support for the different ends and means definition. There is, however, an inherent contradiction between the structure of educational institutions and this idea of equity. So the learning reality rarely approaches the professional rhetoric.

The structure of the school is hierarchical and competitive. The revered qualities are conformity (called cooperation), emotional suppression (called rationality), and acceptance of absurdity (called maturity or respect). None of these qualities is necessary for effective learning. Indeed, all are inimical to learning beyond the trivial. Yet all are necessary for success in learning at a school, because the institutional structure, the political reality that pervades the learning institution, demands these prerequisite responses.

Such an emphasis on control and order is simply incompatible with the idea of young people (of any people) being the main determinants of what they learn and how they learn it. That would be seen by the institution as anarchy. And whilst some teachers would see it as professionally desirable, they would go on to add that "in reality, of course, . . . "

What they mean is that the imperatives of their professional ethic and of their hierarchical morality are different. And in such a situation the hierarchical imperative will hold precedence. Such political expediency is often mis-named "reality". It is more accurately called political obligation, the moral imperative embedded in the institutional power structure. When professional behaviour is not subservient to this obligation, any teacher risks exclusion from the structure. Professional survival is, in the unreal world of the institution, indeed dependent on political expediency.

Equity, frame and hierarchy

Four frames of reference for assessment have been defined; four professionally legitimate ways to describe educational performance, each containing different assumptions about the nature of the task. And each, no doubt, differentially appropriate for particular purposes. Professionally there is an obligation to attach appropriate frames to such particular purposes.

Then three definitions of fairness have been described; three morally justifiable ways to describe educational equity, each fraught with its own limitations, and containing its own implicit notions about the meaning of justice.

These notions of frames and equity come together and form a discourse within educational institutions which are almost invariably hierarchical in their power structures, and these educational systems themselves are embedded in wider societal structures of that very special form of hierarchy called bureaucracy. This is not the time and place to go into detail about differences between simple hierarchies and bureaucracies. At the risk of oversimplification, I will note here that simple hierarchies usually have an identifiable person, with describable characteristics, at the apex. Bureaucracies, on the other hand, are led by shadowy and replaceable functionaries. Personal idiosyncrasies in such functionaries are abhorred. One of their tasks is to await their inevitable replacement by robots with phlegm and aplomb (Arendt, 1969; Kavan, 1985).

Now I want to examine the compatibilities between these professional assessment options, meanings of fairness, and the social structure called hierarchy.

Hierarchy, equity and the Judge's frame

Assessment in the Judge's frame is quite compatible with institutional hierarchy. More than this, by fusing the professional and political aspects of function the assessment process both strengthens and justifies the structure.

Specifically, if the Judge is necessary in order that the student may be accurately assessed, then the hierarchical structure is necessary in order to achieve this educational requirement. In addition, if a Chief Judge is necessary to check, or at least ratify, the accuracy of Lesser Judges, then the next level of hierarchy, the Head of Department, is necessitated. And so on. Thus the illusion that hierarchy is necessary for educational purposes is maintained.

Because the Judge's purpose and power are both based on his or her claim to recognise the standard, the equal treatment definition of equity dovetails nicely with this frame. Indeed, the assessor's work is so much simpler if all students have been through the same educational programme, so all have had an equal opportunity to know or respond to the answers to the questions asked. Whilst Judges would deny the necessity for a rank order of students, they would all be willing to admit that their task is so much easier once the rank order has been produced. All they have to do then is locate the standard between two particular students, and the classification of all the other students automatically follows.

The equal ends definition of equity presents the Judge with no theoretical difficulties. In practice however there are great difficulties.

Whilst the Judges believe they can recognise standards, the research indicates clearly that they are capable only of assessing comparative performance, and the "standard" is inevitably linked to the sample of responses provided, as well as to some assumptions about the composition of that sample. For example, given a large sample with a complete range of student work, a Judge will assess some (or many) as being below the required standard. Later, given a sample containing only those assessed previously at above standard, the Judge will now assess some of these at below standard, especially if he or she assumes the sample is covering the full range (Hartog and Rhodes 1936).

So even if the equal ends definition were achieved with a given group, and through differential treatment they had all reached an adequate standard, according to some data, it is almost certain that the Judge will still assess some at below the required standard.

However, as explained earlier, equal ends doesn't really apply to individuals, but to sub-groups. It's the relative percentage of success between sub-groups that assumes importance for the equity watch dogs. In this regard Judges, being rational and aware beings, are often able to adequately attune their prejudices to the political requirements of their time.

If the equal ends definition of difficulty sets a difficult task for the Judge, then rationally the different ends and means definition presents an incomprehensible one. For how could one hundred completely different products, the outcomes of one hundred different curricula, be compared to a single standard? Surely only Judges of very high status, or extreme arrogance, would attempt such a task.

Faint heart made not fair Judge! To the Judge it's no harder than any other assessment task. The Judge is undeterred by the variety of products and purposes. The Judge's standard is inviolate. The Judge simply compares each work to this standard and the decision is clear.

However, to do this they must of necessity apply their own criteria for success, rather than that of the student. In so doing they would countermand the requirements of an educational program directed towards different ends and means equity, in which the purposes, and hence the appropriate criteria, and thus necessarily the acceptable "standards", vary from student to student. Luckily, such rational considerations rarely impose on the Judge's religious rituals.

Hierarchy, equity and the General frame

The General frame has found little acceptance within educational institutions. Despite the fact that most of the technical and academic literature of educational measurement refers to this frame, and professional testing agencies use this frame for both standardised tests and for grading students, its egalitarian overtones, at least in regard to assessors, has found little response within institutions, despite the overwhelming evidence that using this frame produces more stable rank order grading of students.

Let's look at this a little more closely. The General frame of reference assumes that any single examiner is prone not only to idiosyncratic error due to differences in criteria and "standards" with other assessors, but also to considerable reliability error in his or her own remarking. That is, they will give different marks or grades if they mark the same papers on different occasions, or if they mark different versions of the same paper at the same time. And not only that, but such errors are increased, not decreased, if prior knowledge of the student is available (generalizability errors, that is). And not only that, but that chief examiners are no better than any others in regard to such heinous errors.

All this would be bad enough, interfering as it does with the "right" of the teacher or lecturer to have ownership of their students, and to alone decide their future. But if the assessment input of any competent person is as good as anyone else's, then the whole hierarchical structure of the organisation is called into question.

Worse is to come. Some studies have found that groups of students assessing their own work are also able to get closer to the "true" score than are individual learned superiors. This is democracy run wild; this is destabilisation of hallowed structures; this is anarchy.

Of course, educational institutions can survive without their Judges, although the professional justifications evoked by their presence does wonders for institutional status. If Judges lose the Wars of the Gradings to professional test agencies, then so be it. There are still plenty of hierarchical tasks to be done in selecting syllabuses, administering tests, limiting admission, marking rolls, ejecting students, and so on.

Even so, removing the myth of the Judge from the ideology of the educational institution is pulling out its teeth, leaving it gumless in academia. The function of the school and university has always been equivocal. Rhetorically defined by its purpose of searching for truth and instilling freedom of thought, its practical purpose has been much more mundane - to conserve the culture by perpetuating its myths and reproducing its social and technical elements.

The risk with academics is that they sometimes take their rhetoric seriously, and actively try to bridge the gap between ideology and practise. Given the somewhat radical stance developed in some schools and universities in the sixties and early seventies, it is not altogether surprising that they should be milked of some of their power during the eighties and early nineties of this century. The economic cringe is obvious. But what more Machiavellian way of producing an academic cringe than by using their own research as justification for removing their Judges' power.

In regard to equity, the equal treatment definition implies some measure of competitive merit, and such a measure would certainly be "fairer", that is more stable and less dependent on the vagaries of particular assessors, if the General frame of assessment were used.

This frame would also be useful in relation to the equal ends definition if professionally normed and standardised tests were used as an end point for a satisfactory standard. However, it would be a mistake to believe that the test measured any pre-existing standard. Rather the standard is defined by a certain score on the test. The validity of any such measure is moot. And indeed, this very mootness has left a gap in which the Judge has been resuscitated. For who else is capable to legitimise an arbitrary cut-off? (See any Public examination manual).

The rank ordering procedures of the General frame are not appropriate to the different ends and means idea of equity, because the educational ends and means are individually negotiable, so there is no single "ability" or "trait" or "domain" on the basis of which the students can be ranked.

Hierarchy, equity and the Specific frame

The Specific frame is very compatible with hierarchy. It is the ultimate in accountability and order. Once the outcomes are defined, or the domain of study clearly enunciated, educational programs using computers can reduce the whole educational enterprise to central administrative control, thus bypassing the sometimes difficult professional and technical considerations that in the past have hampered managerial efficiency. New-style managers in particular, wanting clear outcomes and economic accountability, are likely to regard the Specific frame, into which the severely bastardised criterion referenced assessment and competency standards has been incorporated, as a panacea.

Advocates of this frame are likely to down-play, and underestimate, the differences between the equal treatments and equal ends definitions of equity. It's simply a matter of time, they say. Our objectives are clear, our programs are tested, and everyone can reach the desired standard if they try. Some are a little slower than others, that's all, so they will require a little more time. But, given sufficient time, everyone will succeed (Bloom, 1976).

This is facile. Different treatment involves much more than time. Learning styles and appropriate student-teacher relationships cannot be condensed into this single variable. None the less, this could represent some movement towards student empowerment, in as much as very clear and achievable indicators are given to the student about what they must do in order to complete the course adequately.

There is no theoretical reason why some specific behavioural objectives, and some more general criterion referenced objectives, should not be part of the negotiated contracts associated with the different ends and means definition of equity. However, these would generally be negotiated between student and teacher as part of the learning process, rather that imposed on students and teachers as predefined parts of the course.

In terms of its current usage in education, such negotiation would violate current practice and trends, which uses the criterion referenced outcomes, professionally developed and applied, as the true measure of achievement standard. Ironically, to the extent that the outcomes are inadequately defined, and thus confused, the gateway to incorporate such outcomes into the broad definition of equity becomes enlarged. That is, the outcomes may become differentially specific by negotiated discourse with particular students.

Because it denies hierarchy, however, this rarely happens. It is discouraging to see an assessment frame which seemed to hold promise for the empowerment of students now being used as an instrument of rigidity and conformity, as another meter to objectify disadvantage and enshrine privilege.

Hierarchy, equity and the Responsive frame

The Responsive frame contradicts hierarchy. Genuine negotiation implies symmetry of power relations. Openness in communication, the free flow of information in both directions, is not compatible with authority-subordinate power relations. This would be true even if the power relations were reversed, and the student were to employ the tutor to teach. Dependency invariably inhibits truthfulness.

The Responsive frame is also contradictory both to the equal treatment and the equal ends definitions of equity. Responding to individuals in different ways is obviously not compatible with the equal treatment definition, and spontaneous generation of criteria, negotiated curricula and assessment descriptions, and obviously subjective responses, have little connection to common goals and end points.

This is not to say that some well-defined objectives might not be found acceptable and useful to particular students in describing what they wish to learn, and how they will know when they've learnt it. Nor that some other objectives may be so essential to a course that they are prescribed and proscribed in the beginning.

On the other hand the Responsive frame of assessment is quite compatible with the different ends and means definition of equity. This frame is, in fact, a necessary part of any educational processes that value diversity and freedom of students, and thus include this broad equity concept of fairness and justice.

Summary

The relationship of value to assessment mode becomes apparent. Certain definitions of equity, and certain assessment modes, are inherently contradictory to each other and to the power structures that contain them; as such, they will be seen, accurately and probably unconsciously, as potentially destabilising, and consequently be ignored, nullified, or corrupted into acceptability.

In the next chapter we look at the criteria of measuring instruments, and how these fit with the four frames for assessment.

Chapter 9: Instrumentation

Introduction

Assessments in the Responsive mode do not necessarily involve standards or measures. In this frame, assessors may be content to describe without measuring, to give feedback without judgment, to respond with blatant subjectivity.

However, in the political and technocratic world in which evaluation thrives, such 'soft' assessments are scorned, and the claim to measure, to rank, and to compare to a standard is what gives status and power to the evaluation process. Sydenham (1979) points out that even in the physical sciences

> a great deal of modern instrumentation is used to control, rather than gain, new knowledge in the scientific sense. . . it would seem that man seeks to extend the body of knowledge to make eventual use of it to subjugate his environment to suit man's needs (p. 30 - 34).

In the social world, it is people, regardless of any particular label, who are subjugated.

Measurements in physics

To measure any quantity or quality in the physical world we use an instrument, and the instrument must be calibrated. To measure length we need a ruler, and on the ruler is the scale. To measure time we need a clock, and on the clock face is the scale in seconds. To measure current we need an ammeter, calibrated in amperes. The electricity meter measures electrical energy consumed and is calibrated in kilowatt hours.

To calibrate the instrument there are three requirements. The first relates to scale, the second to replicability, and the third to theory-practice bridging.

Whilst scales do not have to be linear (they may be logarithmic or indeed of any other mathematical or ordered function), the nature of the scale does need to be known if any sensible interpretation of the scale is to be made. I will discuss only linear scales here, as they are the simplest and the most common, keeping in mind that the general argument would apply to any other scale for which a mathematical function applies with which to interpret differences.

For a linear scale equal gaps represent equal quantities of the thing being measured. The gap between 3m and 4m is exactly the same as the gap between 6m and 7m. The period of time represented between 9.1 sec and 9.2 sec on the stop watch is identical to the period represented between readings of 12.8 sec and 12.9 sec. The 5 kw hr of electrical energy represented by the difference in meter readings or 39.4 and 44.4, is identical to the 5kw hr of electrical energy represented when the meter reading goes from 44.4 to 49.4. As we pay for the electrical energy that we use, we would want to be sure that this equation was true. We would want to be sure that equal differences on the scale equated to equal differences in energy consumption. And when measures are added we would want to be sure that the laws of arithmetic applied.

We would also want to be assured that our meter gave the same reading as any other meter. It wouldn't need to look the same, or even be constructed the same, but we would want to be certain that if other people used up the same amount of electrical energy that we did, their meters would also indicate that 5 kw hr had been used. So other meters and other occasions must give identical differences for the same energy consumption. Yesterday's 5 kw hr on one meter must be identical to tomorrow's 5 kW hr on another meter.

And finally, after being convinced that the scale was calibrated accurately and the results were replicable, we would want to be assured that the meter really was measuring electrical energy in the units described. We would not want to pay for 5 kW hr of electrical energy if we were only using three. If all the meters are over-reading we are all being equally ripped off, but we are still being ripped off.

To ensure this accuracy we would require comparison with some standard instrument, against which all others could be compared. Such a standard instrument would itself incorporate both the meaning and the value of the thing we are measuring. That is, the standard includes within its operation both the theory of its definition and the practice of its measurement. For example, a standard metre rule is both a practical measure of a metre, and incorporates the theory that equal distances along its length are of equal value. A standard Ammeter, designed to measure electrical current, incorporates within its operation both the numerical value of current marked on its scale, and, within its mechanism, the definition of the ampere as a particular force acting between two conductors a certain distance apart carrying electrical current. And our kilowatt-hour meter gives us a reading on the scale, and incorporates into its mechanism the definition of electrical power as the product of voltage, amperage, and time.

Strictly speaking, such instruments (as instruments), incorporate sub-standards rather than Standards; that is, because they are instruments, they necessarily incorporate an error, which in the cases cited is very small. Because the Standard, which is some fixed point on the scale, is by definition error free, it follows that the Standard must be defined in terms of some mathematical theory (or some replicable event that is more accurately measured than the instrument). That is, with theory or events which have been empirically shown to have specific linkages with other measurable aspects of the physical world.

The standard and the measure

At this point it seems important to clarify the fundamental difference between any standard, and the measurement of that standard, for it is in the failure to appreciate this fundamental distinction that much of the confusion (and manipulation and mis-information) about the measurement of human 'ability' and 'standards' is rooted.

The standard is arbitrary, and is completely accurate. It is not arbitrary in the sense that it is capricious or random. It is arbitrary in the sense that it is based on opinion, and is merely one of a very large number of standards that could have been chosen. However, once the standard is defined as the standard, then it is that exact value. The value of the standard measure is completely accurate not because it has been measured completely accurately; the value of the standard measure is completely accurate because it is a definition, and not a measurement (Sydenham, 1979, p26).

If now we wish to measure a particular thing, we may ask whether it is above or below the standard measure, and by how much. In order to do that we must measure it with an instrument of some kind, or make calculations that involve such measurements. And such measurements will always contain some error, for such is the nature of measurement, because measurements are made along a continuum, unlike counting, which occurs in discrete leaps. We may count the number of bricks, and may do this without error. But no two bricks will be of exactly the same weight. One will have a few more grains of sand or clay than another. And even if two were of exactly the same weight, we could never know that, for the instrument with which we weigh them also contains errors in its scale, in the calibration of that scale, and in the reading of the value of the scale. Two bricks for which we obtained equal weights could indeed be of different weights if measured on another scale of equal accuracy. And two bricks for which we obtained different weights could indeed be the same (within the order of accuracy of that measuring instrument) if measured on a scale of greater accuracy.

One of the party tricks used by educators and others who wish to defend their indefensible measurements is to give examples that reduce measurements to counting. Surely 18 out of 20 correct spelling is 80 percent! Surely number facts in addition or multiplication are either right or wrong! And then they stop. For in the whole field of education they can't think of any other examples where measurement may be so reduced to a counting procedure. Not to mention the sidestepping of the question, eighty percent of what?

The case of the digital watch

Increasingly, instruments use digital electronic mechanisms which use counting methods to give their scale readings. However, these jump from one number to the next, just as watches with visual dials jump forward in one second or tenth of second leaps. Time, however, does not jump forward in such leaps, but is measured on a continuum, as are most of the other quantities that we measure. So the upper limit of accuracy of such an instrument is the gap represented by the jump. The lower limit is much greater.

The interference effect

It is a truism of science, often conveniently forgotten, that any measuring instrument distorts the field it is intended to measure. This is obvious when we think about it. For the measuring instrument to operate, it has to interact - that is, interfere - with the field it is measuring. Newton's Third Law is a universal principal: every action has an equal and opposite reaction; if the field acts on the measuring instrument, then the measuring instrument simultaneously acts on the field.

The effect may be relatively small - a thermometer inserted into a large container of hot water will not much affect the temperature of the water, though it will affect it. However, a very cold thermometer inserted into a very small cup of warm water may cause the temperature to drop appreciatively. The temperature thus measured is not that of the hot water, but that of the water-thermometer system.

In this particular case, it is possible to estimate the imprecision caused by the measuring instrument, if we know the masses and specific heats of water and container and mercury and glass, and the temperature of the surrounding air and the time taken for the thermometer to give its highest reading and the rate of heat loss from the container. Then we may estimate the temperature of the water at the moment the thermometer was inserted. However, even in this simple case, it is necessary to use a theory that is itself, of necessity, subject to some imprecision.

Sometimes the instrument is permanently incorporated into the system, and can then be defined as part of the field. Our electricity meter is a case in point. It is a permanent part of the electrical fixtures in the home. Nevertheless, it does use up energy in its very operation, thus increasing the energy needed for the house. It does distort the field. And as we might expect, it is the consumer, and not the electricity company, who pays for the distortion.

So how big is the interference effect when a 'test' is used to measure some human 'attainment' or 'ability'? How precise is the theory that links the measuring instrument to the thing it is supposedly measuring? And does the test introduce a small distortion into the field it is supposedly measuring, or is it of the same order of magnitude as the field? Are we putting a warm thermometer into the ocean, or into a little test tube of cold water?

Boundary conditions

Another fact of Science often conveniently forgotten is that the precision of the physical sciences - that is, their ability to obtain (almost) identical results in replicated experiments - is directly related to our ability to control the boundary conditions of the experiment: to prevent heat loss, to create a vacuum, to maintain a constant magnetic field, and so on. The precision of physics is specifically related to our ability to create a completely controlled (and hence artificial) environment in which to construct and conduct the experiment. The formulas of dynamics are very accurate in predicting the velocities of objects in free fall in a known gravity field in a vacuum. They are hopeless in predicting such velocities for a skydiver who jumps from a real aircraft in a real atmosphere. She will not reach the ground at the same time as a bunch of feathers or a lead ball thrown out at the same time, nor, luckily for her, at any time predicted by the formulas of simple dynamics. The point to note is that controlling the boundary conditions often produces an artificial environment which makes the data unusable in the 'uncontrolled' world.

This excursion into elementary physics is occasioned not only by nostalgia, but by a desire to clarify some of the relationships between instrument precision and measurement precision in that most precise of sciences, and to point out that whilst precision in Physics certainly cannot be greater than that of the measuring instrument, and any calculation based on that measurement is limited by the empirical accuracy of the attendant theory, that in most cases these two variables are not the main limitation on replicable accuracy. It is rather the stability of boundary conditions, the physical scientist's ability to artificially freeze all other significant variables, that allows such precision, predicability and control in these sciences.

And this is the precise problem we face when we try to measure people. For the boundary condition for stable human behaviour (and all measurement of people, all assessments, all tests, all examinations, must elicit or refer to some form of behaviour), is a stable human mind. But the individual human organism is not a computer. It does not produce a unique response to the same situation, if for no other reason that the 'same' situation never reoccurs. Perception and conception, and hence response, to 'identical' situations invariably differ, as the variables that affect such reactions - attention, mood, focus, metabolic rate, tiredness, visualisations, imagination, memory, habit, divergence, growth etc. - come into play.

As Kyberg (1984) describes it:

> measurement makes sense only when the standards are reproducible, permanence over time being considered a form of reproducibility. Further more, the usefulness of measuring according to this scale depends on some form of reproducibility or permanence among the objects or processes being measured. (p190).

So the very concept of a 'true' measurement resides in the assumption of a stability and permanence in the characteristic being measured, and the boundary conditions of the measurement. Lack of these conditions does not represent so much an error of measurement, as a discrepancy with fundamental assumptions.

Where does the data come from?

Before dealing in more detail with the specific problems in measuring human ability, there is one more point to clarify. Where does the data come from? Where does it belong?

> Data are not out there; they are events interpreted. What constitutes data and what constitutes garbage depends upon frame of reference, aim and method. Furthermore, data are not collected, they are constructed. Data require interpretation and represent the results of a construal, not simply a discovery (Eisner, 1990, p 183).

What Eisner is saying here is very important. The data, the measures, are not out there in the object being measured. They are measures that we have generated through a particular mechanism that includes the measuring instrument and the theory and some aspect or property of the thing being measured. Any claim to 'scientific' truth involves a further implication that a similar mechanism would produce similar data on another occasion with the same person. Or more accurately, with the person that person has now become.

So the temperature is not only some aspect of the object being measured; it is also and equally a meaning generated by a certain way of construing the world (the theory), and a certain way of interacting with it (the mechanism which includes certain actions with instrument and object). As Pawson (1989) expresses it, the only alternative is "to retain the notion of an observable realm that is independent of us yet knowable, . . . (and) to propose some automatic, pre-established harmony between subject, language and world"(p 61).

In like manner, if we are able to measure some aspect of a person called their ability, we are not measuring something they have. We are generating data that is also determined by the mechanism of the instrument - person interaction, as well as by a certain way we, the assessors, have of construing the world. In other words, we ask them to live in our little experimental world for a time, and make a measure in that world. To pin the label on them apart from that world is to misrepresent the experiment: The data, the label, belongs not to them, but to the whole theory-experiment-instrument-object interaction.

Measuring human ability

The rather detailed account of the properties that measuring instruments must have if they are to be usefully used in the study of the physical world enables us to look more adequately at the measurements being used in the study of human ability or human attainment. We might expect such instruments also to incorporate the three same necessary elements: a generally acceptable theory that enables the gap between theory and practical measurement to be bridged, in which a standard measure is defined; an instrument that is itself replicable in terms of the theory, and gives replicable results when measuring the same thing on different occasions; and a scale on which equal differences either represent equal 'ability' differences, or can be translated into some meaningful comparison by a known mathematical relationship. This last becomes particularly important if we wish to use it to make a categorisation, or be added to some other measure.

Standards and standards

Before examining how the Judge, General, Specific and Responsive frames for assessment stand up in relation to these three elements, I want to clarify the meaning of the word 'standard' in relation to human products. This 'standard' relates to a point on a scale, to a point below which the product is unacceptable. The standard thus indicates the lowest limit of acceptability. It requires a scale to define it.

This 'standard' is utterly different to the 'Standard' which is the basis of the scale, and hence of the measures made by the scale. This 'Standard' defines a difference between points on the scale, and can be used therefore to check the replicability of instruments. So we have a 'Standard' metre length, a 'Standard' second of time, a 'Standard' kilogram of mass. I have (arbitrarily) differentiated this Standard with a capital S. Such Standards are useless unless measuring instruments of great accuracy are available to sub-divide and expand the scale embedded in the Standard. However, the specification of any Standard does not guarantee the existence of a suitable measuring instrument (Sydenham, 1979, p 26).

The tendency we have to attribute guilt by association is well known. We are less wary of the tendency to attribute innocence by association. Our Standards of length and time are immensely accurate, as any Standard that defines a scale must be. Indeed, Standards of this sort are infinitely accurate because they are definitions and not measurements. The sub-Standards do involve measurement. And as the sub-Standards also provide bases for scales, the measurements they make must be very accurate and precise. We tend to associate similar accuracy of measurement to those quite different 'standards' that are used to describe minimum acceptability.

Most industry product 'standards' of minimum acceptability are based on criteria for which very accurate measurements can be made. That is, we can measure very accurately whether our product is minutely above or below the stated standard. And that tends to make us forget that the standard itself is not a measurement but is a definition, and is arbitrary. Any amount of a particular additive to food could be harmful to a particular person. All exposure to radiation, even background radiation, has an effect on living organisms. Any bridge will collapse under some particular conditions. Product standards are always statements about a compromise. They represent the arbitrary point at which safety, conservation, style, cost, expediency and whatever strike an uneasy, indeterminate, and hence arbitrary balance. At which point they assume a solidity and stability that denies and contradicts their genesis.

Any standard of acceptability is a political entity, as much in its production as in its enforcement. The myth of certainty that surrounds measures of people is achieved partly by its association with the Standard that defines accurate scales, and with the standard that is a definition of acceptability. As well as the standard we salute as the symbol of authority, as referred to in chapter 6.

Judge's frame

Whilst the Judge often uses a student's written work, in assignment or tests, as a basis for measurement, the Judge would not see the test as an instrument. Nor would he claim to make a measurement. What is written is merely a vehicle for showing him what the student is capable of. The Judge would claim to be able to use any such example as a basis for indicating the level that the student had attained. The Judge is not even particularly concerned to have a sample, random or otherwise. Any example, according to the Judge, can be judged according to its relation to the standard.

In scientific terms, the cognition of the Judge is the instrument, and incorporates the Standard, the scale, the theory-practice gap, the standard of acceptability, as well as the actual measurement, all within its own internal mechanism. Putting it more bluntly, the Judge simply does not operate on a scientific paradigm. Rather the Judge is a mystic who claims to 'know' the definition of standard, rather as one may 'know' the presence of God. A student's level of attainment may then also be 'known' and hence judged accurately, through the union of his/her own consciousness and that of the person being assessed, the example of the work judged being the medium through which this communion occurs, rather in the manner in which tea-leaves activate the astral consciousness of the psychic. Such a process is sometimes conceptualised and rationalised by considering the permutations of such value imponderables as style and form, understanding and creativity, texture and design, understanding of the field, or whatever. Many, if not most judges, would admit however that such variables were used to justify their intuitive judgments, rather than to logically develop their proofs.

From the point of view of the scientific paradigm, the work of the Judge is aesthetic rather than scientific. As such, it belongs logically to the Responsive frame with all the limits and advantages of the overt subjectivity of that frame. Creative reflections on their work by others can be of great value to a student's learning. However, when given in the form of absolute judgments rather than helpful feedback, such reflections are more likely to stifle learning than to expand it, more likely to inhibit creativity than encourage it, more directed to conformity than diversity.

What stops such classification into the Responsive frame is the refusal of the Judge to admit such idiosyncratic subjectivity, and to insist on the truth and objectivity of his judgments as measures of human performance or ability, by invoking the ideology of the absolute standard and the expert judge, and assuming, in both senses of that word, a state of mystical communion.

More recent post-modern conceptions of the Judge's frame use the notion of the interpretative community to defend the position of the Judge. Here quality is determined by a discourse embedded in the language of the field, and various criteria or aspects of quality may be so discussed. However, despite the acceptance within the community of the ephemeralness of the notions it produces, the end result is still the categorisation of the product and/or the student; a solid dichotomous categorisation that denies the tentativeness of its genesis, and, certainly outside that community, and I suspect also within it, is not regarded as a problematic (Fish, 1980).

General frame

The General frame pays considerable attention to problems of scale and replicability, and the theory-practice gap. Theoretically (though almost never in practice) it uses random sampling theory and practice, and assumptions about the distribution of attainment, to produce an instrument (a test), define a scale (normalised score), and estimate replicability (standard error or correlation). In terms of 'ability' measures various standards can also be defined in this model to comprise certain grade levels, in terms of percentiles of defined populations.

Now this is more or less what 'standardised' tests do. In my view they vastly underestimate the error, both in its theoretical definition, as well as in its representation (or more accurately its non-representation) to student and faculty. Some specific details of this are given in Chapter 15 on the psychometric fudge. Rarely do the instruments satisfy the requirements of theory (random selection of items), nor do the populations on the basis of which they are calibrated (random selection of the population). Even so, they do tend to satisfy some of general requirements for a measuring instrument, as required by the physical sciences, even though the errors in these instruments, if made explicit in public knowledge, would make them useless for the purpose for which they are designed.

There are, however, three more fundamental sticky points, points at which the whole exercise becomes very suspect, or unrealistic. The first is inbuilt, and concerns the assumption about normal distribution of performance (or indeed any other assumption that might replace it) built into the theory. There is absolutely no reason to believe that in any area of educational activity the end point should be represented by a normal distribution (which is the same shape as a random distribution) of attainment. In fact, the better the educational environment, the more likely we are to obtain a very skewed, lop-sided, distribution of attainment.

The second occurs when the scores, which are defined in terms of the distribution, are presumed to relate to some 'standard of competence' for an individual student. This latter represents an error in logical typing, but might be more truthfully described as a semantic confidence trick.

Perhaps the most blatant example of this is the distribution grades that are labelled A B C D F. These grades may be defined in terms of percentile distributions, so that A represents the top 5 percent of the rank order of students (or whatever other arbitrary percentage is chosen), B the next 20 percent, and so on. Logically then, F represents the last 5 or 10 percent or whatever. So why not E? Because F also stands for 'fail', a statement about competence and not distribution. And historically, as we know, A and B have connotations of excellence that C does not have, though there is nothing in the distribution that implies either that A is an excellent performance, nor that C is a mediocre performance. For example, if a group of professional sprinters throw the javelin and are then graded in terms of their rank order, we would not expect those obtaining an A to have reached the Olympic 'standard'. On the other hand the person who runs last in the Olympic 800 metres final is hardly a mediocre runner, or a failure.

For even if we except the notion of a 'normal' distribution, the sticky question still remains: a normal distribution of which group? All the people in the world? All the educated people? All the people still at school? All the fifty year olds? All the people at a particular grade level? In a school? In a city? In a country? Without this detailed information the 'standards' cannot be given a meaning. And even with them, they can be given no meaning other than that defined for them. That is, their meanings can only relate to distribution, and not to competence.

Even with such information about the nature of the sample population, there is, and can be, no formula, no equation of equivalence, between grades defined by distribution on a rank order, and some pre-specified level of attainment of an individual student (Airasian, 1979, p 42; Jaeger, 1980, p 64; Glass, 1978; Levin, 1978, p 314; Burton, 1978, p 263; etc.).

In addition, the differences in logical type in attempting to make linear measures of complex qualities generate paradox and confusion and hence strong emotion and unresolvable debate (See Chapter 12). This makes the topic utterly suitable for creative endeavour and satirical humour, but impossible for scientific measurement.

The third point is more fundamental, and may well make the other two points trivial. There is no Standard against which the scale can be calibrated. There is no theory that enables a definition of some point on the scale to be distinguished, against which the scale might be calibrated, along with other scales purporting to measure the same thing. The test scale floats freely in space, relating solely to its own assumptions with no Standard rope to bind it to the earth. What we have here is not a scientific instrument, but a very suspect ordinal scale pretending to derive from a scientific measurement.

Specific frame

In the 'pure' Specific frame, a person's 'ability' or 'performance' or 'attainment' is reduced to a finite number of specific behaviours, for each of which a 'standard' is clearly defined. Thus we are, in theory, able to specify exactly which 'objectives' have been achieved to the specified 'standard'. The notion of scale, Standard, and measuring instrument is (apparently) sidestepped by postulating a dichotomous variable, requiring not a scale, but rather an on-off switch, to categorise its measure. We shall come back to this in Chapter 11, where it is argued that all categorisations infer measurements.

However, in most areas of human endeavour such reductionism to specific behaviours results in trivialisation of the task. Further, specification of the 'standard', even in such a narrow and specific thing, is still very difficult in most cases, as the measurement instrument does not exist, and it is finally fallible human judgment which in practice must decide whether the standard has been achieved for each objective. Further, the basic assumption is erroneous; the variable being measured is continuous, not dichotomous, so the measurement error still exists, disguised though it might be. We are back again to the Responsive frame, requiring a subjective decision, which is covered up by pretending to be the Judge's frame, requiring an unambiguous omnipotent objective decision, which is in turn covered up by pretending to be an example of unambiguous standard in the Specific frame, derived from a definition of standard which pretends to be dichotomous and pretends to be nonarbitrary.

To further confuse the issue, what often now happens is that specific information about which particular objectives have been achieved is lost when measurement is reduced to counting, and the number of objectives achieved is the only information recorded. This creates the illusion of exactness and error-free information by disguising the fact that the exactness of the 'standards' of individual objectives is, in practice, illusory.

Responsive frame

In the Responsive frame the person's work, or inferences about the person's capacity or ability, are described but not measured. Further, these responses are ideally owned by the responder, and not projected onto the producer, or the producer's work. They may describe how the person's performance relates to certain criteria, how then the performance might be improved, and to what extent, in terms of such criteria, and in the opinion of the responded, success has been achieved.

The responded may also offer some opinion about whether the work of the person being assessed is of inferior or superior quality, or whether they are skilled enough to practice in a certain field of work. However, again, this does not purport to be a measurement of some clearly defined standard, but merely the informed view of a particular person who for some reason or another has views worthy of hearing. As Stake describes it:

> People do not just disagree, they live in different realities. People live quietly and often proudly with their peculiar ways of seeing things. The evaluator errs in too noisily depicting the peculiarities as much as too quietly. . . . multiple views help legitimate resistance to bureaucratic standardization. (Stake, 1991, p 85).

But note how quickly Stake modifies the insight of his first sentence with the caution of the second. In whose interest is this emphasis on quietness? Why this concern to legitimate resistance rather then stridently call for reform? Who might hear strident voices, that quieter ones may not discern? And whose voices go unheard in the quest for quality, and the demand for categorisation?

And note also the very narrow gap between offering an opinion on whether the performance is adequate for some purpose, and categorising the student. We are here at the very edge of the Judge's frame of reference, a boundary crossed over as soon as the categorisation is made.

Summary

In this chapter we have looked at the invariances required in events involving measuring instruments if such events are to have credibility. In particular the notion of a Standard that theoretically defines the scale, and how that is not to be confused with a standard of acceptability, which is to be measured by the instrument, and which requires a scale in order to be located. We also noted the importance of the specification of boundary conditions and interference effects, and that the price of invariance and tight theory-practice links was artificiality.

The various assessment modes were then analysed in terms of their instrumental error. All were found to be invalid, on the grounds of not satisfying the conditions of adequate instrumentation.

Chapter 10: Comparability

Synopsis

In this Chapter I examine the notion of comparability as it applies to the assessment process. Any rank ordering of students, any adding of marks on examinations, any addition across subjects, assumes that comparisons can indeed be made.

The fundamental distinction between more and less, and better and worse, is first elucidated, and this is linked with ideas of uni- and multi- dimensionality and notions of doing or having. This analysis is then applied to ideas of traits, abilities, and skills, and their supposed measurement in tests and examinations. Some fundamental confusions are exposed.

The discussion then moves to what meaning if any can be given to the result when marks or grades are added, how loadings on final rank orders are affected by spread of marks, and how differential privileging of sub-groups occurs with different intercorrelations. Finally, it is contended that for individual students the privileging is non-predicable, and the total score thus meaningless.

Goal kicking skills

George!

Yes coach?

You know why we've lost the last six games?

The other teams were better?

Bad kicking, George. Bad kicking. And with six in a row, someone's got to go.

Gee coach, that's really poetic.

Yeah George, and you're really pathetic. Anyway, do some tests and get me a team ranking on best to worst on goal kicking skill.

No worries, coach. Goal kicking skills, you said?

That's what I said. Get me a best to worst ranking on goal kicking.

What particular aspects of goal kicking, coach?

You're the trainer, George. How far they can kick. How straight they can kick.

Anything else?

Jeez, what do I pay you for? Set kicks, kicks on the run, and snaps. That ought to do for a start.

No worries, coach. I'll work out some tests for each of those and give you a list in a coupla days.

(Two days later).

Here you are, Coach. Here's the list. I've ranked twenty five of them in order of merit on goal kicking skills.

That's great, George. Just what I wanted. Let's have a look at this. Harvey's on top of the list. How many goals has he kicked this season?

None, coach. He's been playing in the back pocket.

Look where you've got Shonker. Twentieth. He's the bloody full forward. He's booted a hundred goals this season already.

Yeah! well, he's missed two hundred.

So he's missed two hundred. He's still booted four times as many as anyone else.

That's because he has ten times as many possessions as anyone else. You didn't ask me about that. You just asked me about goal kicking skills.

Yeah, OK. So who's the longest kick?

Can't tell you that. It got lost in the data.

Who's the most accurate on set shots over 50 metres?

Got lost in the data.

Who's the best snap shooter. No, don't tell me. Got lost in the data.

Hate to tell you, coach, but I think this list is a load of shit.

You can say that again. Who was the idiot who did it?

The idiot who did what some other idiot told him to do.

Better or more?

Fundamental to the process of arranging orders of merit is the notion of comparability. As we have seen, the notion of standard implies the notion of order of merit, which implies the notion of more or less, better or worse. For such notions to have a meaning, they must refer to some aspect, some property that is being compared, that is presumably being measured.

Regardless, the first paragraph slid past a fundamental distinction: "more or less" is not the same as "better or worse": More or less are terms related to counting, to mathematics, to scales and measurements. They are loaded with notions of objectivity, and solicit entry to the quantitative world; better or worse are terms related to value, to goodness. They are permeated with the aura of subjectivity, and are related to the qualitative world, the world of valuing. The concepts are in different domains of discourse. If the criteria is size, then two people may be compared as being more or less heavy; or their weights may be compared in terms of better or worse in regard to health. But the two ratings are unrelated. Or if the criteria is emotionality, we may rate people in terms of whether they are more or less emotional; or we may rate them in terms of the appropriateness or productiveness or empathic clarity of their emotionality. Again the two ratings are conceptually unrelated. Or so it would seem.

What is the essence of this difference? For when we tried to explain what we meant by better, we used words like healthy, productive, empathic, clarity: and the interesting thing is that we may use more or less with any of these words, even though we started off in the better or worse category. And we may also ask of each of these new criteria whether they are better or worse; in this case questions preempted in the predominant paradigm because value judgments of better are already built into the words chosen to describe the criteria.

So what is the essence of the difference? In relation to aspects like size or emotion or clarity, when we ask the question more or less we are asking about intensity, about how much or how many. We are referring to the aspect in isolation from its environment. The event that produces the judgment about more or less involves our sensory relation to that aspect independent of other aspects. More or less questions are answered by focussing on the aspect and on no others. More or less questions are directly answerable. The answer may be incorrect, but such a statement in itself implies that there is a correct answer. More or less has only one meaning in relation to a particular aspect. They can't be more and less at the same time, so the question is convergent, and presupposes a world in which there is a true answer to the question. So logically more or less implies a uni-dimensional aspect, a world of transitive and asymmetric relations (Lorge, 1951, p548).

On the other hand, when we ask the question better or worse, we have to ask another question, In what way better or worse? Because something may be better in some ways and worse in others. Better or worse in what aspects? Or better according to whom? Or better under what conditions? And when we nominate those aspects we can ask of them two questions about any comparison; more or less, or better or worse. And so on. Essentially better or worse implies multi-dimensionality in the aspect under consideration.

What does all this mean? Very simply, when we ask the question more or less there are no further questions to ask. We move straight on to the answer. In other words, more or less questions define the end of discourse; they are a direct invitation to a judgment; they are the signal to stop thinking, and act; and incidentally and significantly, to accept the judgment, which comes after the thinking has stopped.

But the question better or worse logically invites more questions about the first criteria. In what way better or worse? Which introduces more aspects, particular aspects selected in most cases from a much larger set of possibilities. For there are as many aspects as our conceptual imagination may produce (Lorge, 1951, p536). Yet the original aspect is reduced, even as more precision is generated by defining aspects; and as more aspects are conceived, the potential disparities of the judgments concerning them increase. And then for each of those aspects: More or less? Better or worse? And again, the additional questions about positioning and context are generated. So better or worse questions encourage further discourse, and further thought.

All this is not to deny that the power relations in which such discourse is embedded may dictate that the answer to the question better or worse be given at any time and be accepted without further thought. But that in no way invalidates the additional logical questions that the aspect implicitly generates.

Having and doing and being

It is obvious, but important, to make the point that whole entities (holons) cannot be directly compared in terms of more or less, only aspects of them (Jones, 1971, p335). One dog cannot be more than another dog. Nor can a stone be more than another stone, nor a stone be more than a dog.

In like manner dogs and stones cannot logically be compared in terms of better or worse, for such a claim is meaningless without a response to the question "in what way better?" A dog cannot be better than another dog. In terms of dogginess, dogs are equally doggy; they are equal by definition, as being classified as dogs. Likewise with stones. And dogs and stones cannot be compared as entities because they are in different classes. It follows that the very act of classifying whole entities (into classes) logically invalidates any comparisons within or between the entities that comprise them. Classes of course can be compared in terms of the numbers of elements they contain, but this is a different matter.

Two people are being compared in terms of the relative merit of some task. In terms of doing, we may say that one person does it better than the other. This is a statement about relative merit. Or we may say that one person does it more than the other. This is a statement about relative frequency, and not of relative merit. You may drive a car badly many times.

In terms of having, we may say that one person has more of something than the other. This may claim to account for the greater merit. It is essentially a statement about the comparative number of elements in a class. But we would not account for a difference in merit by saying that one person had that something better than the other. Such a statement refers to the whole class and whole classes cannot be compared except by numbers of elements.

So in terms of relative merit, the question of more implies a different mode of description, a different ontology, than does the question of better: Better or worse is a comparison of what people do under certain conditions, made by some person; more or less is a comparison of what people have, or are alleged to have. As such it is logically independent of any contextual or positioning variables. One begins to see the simplistic delusion generated by mathematical modelling.

Logically then better or worse questions cannot be answered definitively until they are reduced to a criteria which comprises a class in which the question better or worse is reduced to the question more or less. Logical here means relations that are transitive and asymmetric.

Pragmatically, better or worse questions can be answered whenever the criteria are sufficiently understood (implicitly or explicitly) to allow consensual subjectivities of judges to give similar answers. However, as we have indicated earlier, such criteria are multi-dimensional. And as is evident from the conversation that began this chapter, little if any meaning can be given to a uni-dimensional description of this multi-dimensional entity in terms of their uni-dimensional elements. As we shall see later, one meaning of such a comparison is dependent on the relative loadings of the different dimensions.

Politically, of course, better or worse questions are answered whenever someone with sufficient status or power gives a decision.

Comparing people

It follows that to compare people, whole people, we may compare either some parts that comprise them, or some wholes of which they are parts. If we look at the parts that comprise them, we may look at the person's elements or internal processes; if we look at the wholes of which they are parts, we may examine the person's functions and relations in the wider environment or community, or at the cultural meanings in which their thoughts and actions are embedded (Wilbur, 1996).

Let us compare two people in terms of their relative merit in Physics. We are particularly interested in their relative achievement in a particular course of study at year 12 level. Such a course has a range of content and objectives and involves practical and cognitive operations of varying complexities.

We are obviously in a multi-dimensional world, in which at this stage more or less questions are meaningless. Further, any logical answer to the better or worse question is going to depend on the details of the answer to the prior question: In what way better? What particular aspects? Under what particular conditions? In whose opinion?

And if we intend to give a meaning as well as an answer to a multi-dimensional comparison, what are the relative loadings of each aspect in the final judgment?

Of course, we could simply ask the teacher who taught them, who is better? And the teacher might give a judgment. But in making sense of that judgment in terms of the original question, the implicit questions still hang there; in what way better? So after the judgment, the teacher must logically justify the decision on the basis of criteria; and if one is not better on all possible criteria, then the question of how the criteria are loaded to obtain the final criteria is relevant.

So, either prior to or after the judgment, how might the discourse progress?

> *In what way is she better?*
>
> *She knows more facts.*
>
> *Is that all?*
>
> *No. she's better at solving problems?*

In what way better?

She gets more complex problems right?

Does she get more simple problems right?

No, he gets more simple problems right?

In what ways is he better?

He is more careful, he makes less mistakes.

And so on , and so on. And if we are dealing with twenty or thirty persons, it is clear that different criteria of comparison are possible for each pair, and there is no reason to believe therefore that there would emerge any final rank order of merit, for on the basis of different criteria of comparison, A could be better than B on criteria 1, B could be better than C on criteria 2, and C could be better than A on criteria 3. This is an empirically inevitable consequence of multi-dimensionality. It is inevitable because only when every criterion correlates unity with every other criteria will ranking invariance occur. And in that situation we are, by definition, in a uni-dimensional situation. It is the reason that psychometricians fantasise unmeasurable but uni-dimensional true scores.

Viewed from this perspective, it becomes clear that the more specific, limited and applicable to all comparisons the criteria become, the more possible it is to finally reduce such aspects to those answerable by more or less, the more possible it is to produce an invariant ranking, and meaning (in terms of explicit loadings) for the meaning of the original comparison. However, such meaning is at the expense of initially reducing and finally confusing the meaning of the original comparison. Another example of the essential contradiction between reliability and validity.

Traits, abilities and skills

A trait or an ability is a thing that a person has. A trait is a hypothetical entity, an abstract attachment, a comparative label, that is used to explain differences in what people do in terms of something that they have. A trait is described not so much as a performance as a potential performance, as a sort of template of the performance that might emerge under ideal conditions, whatever that may mean; a morphic field that predates performance. This magical property of a trait makes it forever immune to particular environmental conditions, which may indeed influence particular performances, but leave the trait, securely protected within the person, unsullied and unmoved, firmly fixing individual merit in correct relative position in the grand order of things.

A skill is a much more difficult ball of wool to untangle. A skill is something you have, like a verbal reasoning skill. On the other hand, a skill is normally exhibited as something you do, like playing a musical instrument or tennis. And you can have more skill but maybe not better skill (skill here is used as a holon). On the other hand, you can have more skills or better skills, and these two meanings are different, as with the goal kicking skills referred to earlier. Better skills here appears to have more to do with a particular selection of skills relevant to a particular context. Then again, skill seems to refer at times to a particular standard in a more-less or better-worse ranking; unskilled refers to rankings below the standard. It is clear from all this that the word skill is a very useful word to have in any discourse that wishes to imply precision even whilst it multiplies confusion. Norris (1991) notes a similar confusion in the notion of outcomes:

> The precise specification of performance or outcomes rests on and leads to a mistaken view of both education and knowledge. Mistaken because there is a fundamental contradiction between the autonomy needed to act in the face of change and situational uncertainty and the predictability inherent in the specification of outcomes (p335).

The world of objective tests

Objective tests, which often claim to be value free, necessarily do not ask better or worse questions. The whole operation is contrived so that only more or less questions are asked and answered. Further, they necessarily deal with what people have, not with what they do. Thus it is not so much a desire to deceive that drives the psychometrician to imagine constructs such as ability or traits or skills, but a logical necessity of the world they have constructed.

For it follows that if there is to be an answer, rather than a multitude of answers, to a comparison of two people, it is essential that the question better or worse never be asked, and all comparisons be reduced to the question more or less.

So the world of objectives tests, like the world of chess, and the world of mathematics generally, is certainly internally logical. Whether it relates to anything that actual people do in the world, apart from answering objective tests, or playing chess or mathematics, is another question.

The world of public examinations

Examinations live in far more dangerous territory. The constructors and markers of examinations are far less isolated from the front line of educational activity than are test writers. Their language is less precise, their pragmatism more up -front, their compromises and contradictions more overt. So they are far more likely to slide uneasily between concepts of better or worse, and of more or less, according to the pragmatics of phases of the assessment.

Consider the marking of essays. Whilst guidelines for marking may be given, ultimately notions of better or worse must be utilised by examiners in deciding what mark to give. Such guidelines are designed to circumscribe the answers to the question "what aspects?," to limit variability in the question "who says it's better?," and hopefully bypass entirely the question of the effects of the conditions on the essay's production.

So in stage one, the answer to the question of "better or worse," which establishes the ranking of students on a particular question, is used to determine the answer to the question "more or less," which is the mark given. Now the marks are added to give a total score, which is then interpreted as being better or worse according to whether it is more or less. Finally, if the grades are not distributed statistically, someone must look at whole papers around the grade boundaries to decide which are in their opinion better than the standard that defines the boundary, and which are worse.

Now, it is clear that this procedure only makes sense if the notion of better or worse, and the notion or more or less, are synonymous, within the series of events that comprise the examination. In other words, if better means more within the context of the examination. Practically, this makes it now impossible to untangle the interaction between the two notions, or deal with the complexities involved when multi-dimensional aspects are mapped onto uni-dimensional scales.

It is not my intention to suggest a solution. It is my intention to establish a confusion, and to note that such confusions must invariably lead to more invalidity and uncertainly about what is being described here. In other words, here we have another, crucial and fundamental, source of error.

We are tapping here one of the distinctions between quantity and quality, two concepts often fused together in discourse on measurement and evaluation. At this point it is sufficient to note that big is not necessarily better; getting more sums correct than somebody else does not necessarily make you better at mathematics: nor does getting more spellings correct make you better at writing, or getting more multiple choice questions correct on a philosophy test make you better at philosophy, or a better philosopher. To suggest otherwise is perpetrate a category confusion. The matters raised in this paragraph are further elucidated in Chapter 12.

What can be compared? What can be added?

So in terms of "more or less" we can compare any events that have a common aspect, that have a criteria on the basis of which we can rank them in terms of having more or less of that common aspect. A criteria, that is, that can be considered uni-dimensional.

Two questions then arise, which are fundamental to the whole notion of testing, examining and credentialling. The first question is, what happens when we add measures or ranks that relate to the same aspect? The second question is, what happens when we add measures or ranks that relate to different aspects?

Let's compare swimming pools in terms of two aspects that are comparable in terms of the same measurement units, a claim incidentally we could rarely make in the human measurement field; we could compare the pools in terms of length, or in terms of depth. In both cases they may be measured accurately (to within one millimetre) in metres. Now we could obviously compare our pools in terms of length, and we could compare them in terms of depth. The question is, could we use these criteria to obtain a single measure in terms of which they could be compared? This is in many ways an ideal situation; we have an accurate scale and measuring device, and our two aspects can be accurately compared on the same scale. So we could add the measure of length and the measure of depth. But what would it mean?

We could classify swimming pools uni-dimensionally in terms of the sum of their length and their depth. In terms of the initial components we have now lost any meaning, but the process (the addition) does enable us to imply another meaning; in this total positioning length and depth were equally valued, because we added the two measurements together, each with a loading of one. Or so it would simply appear. But things are not always what they seem and in this instance this would be an erroneous inference.

The relative valuing of the two components may be looked at in two ways; in terms of absolute value of the combined measure, or in terms of the influence on the rank order of the combined measure. Let's look at the absolute measures first.

If the depths of the pools varied from 1 metre to 2 metre, whilst the lengths varied from 10 metre to 100 metre, magnitude of the addition would be almost entirely defined by the length measurement. Alternatively, if the lengths of the pools were all between 15 metre and 16 metre, and the depths varied from 1 metre to 5 metre, then again the length would contribute most to the total measure.

However, in the second case the final rank order of the total measures would be most influenced by the depth measurement, which has a bigger range. So whilst the loadings for absolute values of the sum of measures are determined by the absolute values of the components, (which could statistically be characterised by their mean value, if we wanted to lose a lot of information), the loadings for determining the final rank orders are determined by the standard deviations of each component (Guilford, 1965, p424).

In this situation, the rank ordering of the total can be given a (process rather than content) meaning in terms of the relative valuing of the two components; and that valuing is implicitly determined by the standard deviations of their measures. We may adjust this by loading one of the measures. For example, a diver may greatly value depth over length in his pool, so may want the addition to mirror that valuing. So the diver may want to load the depth scores (by multiplying by a certain number) so that the standard deviation of the (loaded) depth measure (before addition), is 5 times that of the length measure. On the other hand, a long distance swimmer may want the two dimensions loaded the other way. In both cases the specific loadings are arbitrary, and in both cases they are related to function. And in both cases the final measure has no meaning other than that attributable to the relative contribution of each component to the final measure. (Of course, in this case the addition was completely unnecessary to the function; it would have been more rational for the diver to specify a minimum depth and minimum length, and for the long distance swimmer to do likewise; but that would have left us with no single variable with which to compare pools. And as mentioned elsewhere in this thesis, that may be the whole point of the exercise).

Let me generalise a little from this very simple case;

- 1. Any measure implies a ranking. Rankings imply transitive and asymmetric relations.
- 2. Rankings of a single aspect have a meaning, in terms of relative size or intensity of that aspect, which we can specify as more or less, and hence by numbers.
- 3. Rankings of different aspects may be added, but the addition has no meaning in terms of either of the aspects taken separately; the addition can be given a meaning in terms of the relative contribution of the two aspects to the total.
- 4. The relative contribution to ranking is determined by the loadings, equal to standard deviation multiplied by an arbitrary number.

The effect of correlations on loading

Let's go back to test and examination scores. We have three sets of scores (L, M, N) for the same group of people. The scores have the same standard deviation. We wish to add them to get a total score. Our theory tells us that they will have equal loadings on the final score.

Assume L and M scores correlate zero. Then when we add the L scores to the M scores, rank orders of both are changed, and it looks as though they contribute equally in determining the final rank order.

Assume M and N scores correlate one. Now when we add the N scores to the M scores the rank order of the M scores is unchanged. We could argue that the N scores have contributed nothing to the rank final order.

But then, if we add the M scores to the N scores, we could argue that the M scores contributed nothing to the rank order. A paradox. It is not necessary to resolve the paradox to realise that in this case the loading is determined by what is being added to.

It is also very clear that the final rank orders are very different in the two cases of zero correlation and unity correlation. Regardless of the loadings (statistically determined by the standard deviations), different students have been privileged in the two situations described. In the uncorrelated (r = 0) groups, no particular group of the M score group is being privileged, or under-privileged, by the addition. However, in the perfectly correlated groups (r=1), the students who do better in M scores are all privileged when the scores are added, and the students who do worse do worser when the scores are added. This is in addition to the fact that the standard deviation of the composite score is 1.4 times greater in the case of the perfectly correlated group, giving it just that much extra loading as a composite when compared to the other total (Guilford, 1965, p418).

So what does all this mean when both L and N scores are added to the M scores to obtain a single rank order? The L and N scores both have equal loadings to the M scores; but this is a group phenomenon, and tells us little about individual students or sub-groups of students. We have seen that the L score loadings are more or less equally distributed across the M scores, but the N scores have privileged the top sub-group (according to M scores) and down-graded (with respect to the total score) the bottom sub-group. By interpolation we can see that this phenomenon will have a differential effect over the whole range of possible correlations and will be greater as the correlation with the scores added to increases.

In addition, to the extent that the means of the L and N scores are different, to that extent will the addition scores generally privilege the group with the higher mean.

It is clear that the statistical notion that relative standard deviations determine loadings is a vast oversimplification when applied to complex comparison situations.

Comparability, true score, and error

Here we have presented, in very simple form, one of the dilemmas of public examiners who must cope with adding different scores, from different subjects, or from the same subject marked internally and externally, and end up with some final rank order of marks because someone has said this is what they must do.

I have argued that such a total score can have no meaning other than that inherent in the loadings attributable to each component added; and I have shown that whilst the loadings of the whole group from any one school may be controlled through controlling the standard deviation of the marks, the correlations of the score with the score added to will influenced the subgroups which are over or under privileged by the addition.

There is another paradox evident in the conclusion, especially in regard to internal-external scores. To expose the paradox two further facts need to known.

Firstly, the rationale for internal assessment is that something different (broader, deeper, more complex, more varied) is measured by the internal assessment. Secondly, we can assume that in most public examinations some twenty to forty percent of students will be deemed to have failed, and to that extent the rank orders of their final scores are irrelevant in respect to the grades of those who pass; so the pragmatic teacher might argue that to underprivilege students who will fail anyway "does not matter."

In such a situation, it is rational (if somewhat inhuman) for schools to aim for maximum correlations with the external examination in order to privilege those who will most benefit from such privilege (that is, the best students). However, in order to do this they must invalidate the internal examination; for such an examination is surely more valid the less it correlates with the external scores, because it is supposed to be measuring something different. In short, the price of success is invalidity.

The middle way

> *That's all very well for the front runners, but most of the kids I teach are more middle of the road. I just want to get as many as possible past the cut-off point for entry to University or TAFE.*
>
> *Well, you've got a different problem then. You want to maximise opportunity for the middle group, not the top group.*
>
> *I suppose you could put it that way. So how do I do that?*
>
> *Easy. Just take out that middle slab of students and put them at the top of the rankings.*
>
> *Just like that?*
>
> *Just like that!*
>
> *But isn't that unethical? Doesn't that make the whole examination invalid?*
>
> *Sure. But as I've explained, it's invalid already because of what many schools are doing for their top students.*
>
> *Are they really aware of what they are doing?*
>
> *What's the difference. I don't accept the view that in this case bliss in ignorance makes the position less unethical. It certainly doesn't make the practice less invalidating, or the errors less significant.*

When equal loadings are unequal

I have shown how equal loadings for a group may take on different shapes according to the correlations. Equal loadings for a group does not in practice mean equal loadings for all subgroups of that group. And in terms of individual students it doesn't have any particular meaning.

The question then arises, does equal loading for the whole group of students mean equal loadings for each separate school? Surely some school groups are really better than other school groups so should be differentially loaded? Some school groups might have higher means, and some may have larger or smaller standard deviations in the sets of marks that indicate their comparative attainments. And these might mirror differences in intrinsic ability, whatever that means, or might be a function of very good, or very bad, teaching, whatever that means. But if such students are tested internally, how would we know about their differential potential, or their differential attainment, as distinct from differential testing effects? And especially how would we know if they study and emphasise different things, and value different criteria, so that their results are essentially non-comparable? Or if they study different subjects, with utterly different realms of discourse, such as chemistry and Japanese?

Now there are a number of ways of trying to solve this problem, all of them more or less inadequate. McGaw (1996) summarises them well: use some external examination (either the specific one related to the subject, a single "scholastic ability" test, or some grand total score on all external examinations) to statistically adjust the internal school results; this is statistical moderation of the school-based assessments. Or alternatively "use some external review and checking of schools" assessment results by teachers from other schools or authorised assessment experts to control the level and distribution of school-based results (ie consensus moderation)" (p82).

Such moderation systems provide different processes for modifying the means and standard deviations of school scores on the basis of comparison with other scores or other schools or other students. To the extent that the correlations with the criteria (whether the criteria are scores or actual criteria in the minds of the moderators) are high, to that extent is the moderation reasonable, and possibly invalid. And to the extent that correlations with the criteria are low, or differential, to that extent is error compounded, as we have indicated in the previous discussion.

I do not intend to enter into the debate as to which of these is the "best" way to go, or indeed whether they all do not produce solutions which are more inequitable than the problem they were devised to solve. My project here is not to indicate how such problems may be best solved, but rather to detail what implications such solutions have for the empirical determination of error.

Comparability error

What is clear is that different solutions, including no solution, produce different results. The notion of "true score" is dependent on the notion of some uni-dimensional trait that is obviously non-admissible when the additions involve not only components which have low correlations and do not claim to be about the same thing, but the different additions contain different components. (That is, different additions contain marks from different subjects) But the notion of difference in estimates requires no such theoretical underpinning. It is empirical data demonstrated by differences in empirical rankings or scores under different experimental conditions.

Estimates of comparability errors are easily computed. Given that various forms of inequity are inherent in all measures of both school based and external examinations; that the meaning of the final rank order is based on relative loadings; that all means of trying to create equal loadings involve the creation of arbitrary assumptions and the subsequent construction of additional inequities. Given these facts it is relatively simple to construct a number of different aggregates according to the various models available (including the original raw data), and thus determine the range of ratings (or scores) that these produce. These empirical differences are an estimate of the comparability error. Such a set of scores has the added advantage that it relates to estimates for each individual, and does not confuse such individual differences with group statistics (such as standard error of the estimate).

Note that this is not the assessment error. The comparability error is the additional error added through the procedures of summating or summarising scores, which are independent of other sources of error described elsewhere.

The ontological remainder

My description of comparability error here begs the question as to whether the whole process isn't a nonsense, because of the meaninglessness of the total score. In order to examine that notion briefly I will examine the construct, not of academic merit, which might be a name that we could give to the sum of marks on test or examination performance in various academic subjects, but rather the idea of athletic merit, a similar construct we might conceive in the field of more physico-social endeavour.

> *Concerned at the physical flabbiness of our youth, the party in power in the Federal Government, as part of its election platform for 1998, promised to improve the nation's health by removing the flab.*
>
> *Thus in the year 2000, two lists of year 12 students were produced by Education Departments in each State. One for academic merit, and one for athletic merit. Students are required to nominate three areas of physical prowess. To ensure some breadth they must include at least one area from athletics or swimming, and one from team sports.*
>
> *Brad and Diana make their choices. Brad, who does not like running, and is not very strong, chose walking as his athletics choice, doubles bowls as his team game, and pistol shooting. Diana chose the hammer throw for athletics, basketball for a team sport, and golf for the third choice. Diana is not very fast or indeed very agile, but she is 1.8 metres tall and weighs 95 kg.*
>
> *Brad and Diana both covered the curricula designed around their choices, and completed the various tests designed to measure their skills in the designated areas. After some statistical corrections, their separate scores were added to give a final mark. They both obtained the same score of 189 points which is about half a standard deviation above the mean for all year 12 students in Australia.*

> *Independently of this (obviously), they were both offered scholarships at the Australian Institute of Sport; Brad because his pistol shooting scores place him in the world's best ten; Diana because last year she broke the Australian Women's open hammer throw record.*

This story is important because it is about individual students and not about groups of students. All of the talk of equal loadings and fairness is in the "equal ends" definition of equity. It attempts to address inequities involving groups of students, but in no way addresses the inequities done to individual students. And just as attempts to address inequities between whole school cohorts invariably leads to other inequities in terms of sub-groups within the school, so any attempts to reduce "better or worse" questions to more or less questions, or any attempt to reduce multi-dimensional entities to uni-dimensional ones, must invariably discriminate against some students more than others, and utterly confuse the meaning of what the final ranking is really about.

The second aspect of the apocryphal story that I want to draw attention to is its obviousness. It is obvious that all of these physical activities are different from each other and that whilst comparisons of aspects within a single sport may sometimes be meaningful, between sports such comparisons are meaningless.

What is not so obvious perhaps is that the complexity and possibilities of difference within cognitive endeavours have much more span, and much more depth, than do those of a largely physical nature. For this field encompasses the whole universe of cultural experience and knowledge. And the ideologies of schooling, if not the practices, assure us that students will have the opportunity to tap this richness. Even so, at the end of the day it all gets reduced to a uni-dimensional list. And both the tragedy and the absurdity of this gets lost in its normality.

Chapter 11: Rank orders and standards

Synopsis

In this chapter the relationship between rank order and standard is teased out in more detail: In particular the particular meanings given to the standard in the Judge and General frames of reference; how logical confusions proliferate when discourse jumps from one frame to the other; and how the differences in meaning are connected logically.

At the end of the chapter a post-modern myth of the situation is presented.

Personal day-dream

I was about fourteen when I first pondered the sticky issue of the elusive standard. The context was heavenly, rather than earthly, theological rather than educational.

It concerned St Peter. It seemed to me he had a problem. Here he is at the pearly gates as the newly dead file by and do their thing - state their case. And Peter, judge extraordinaire, gives his verdict; pass, fail, pass, fail, fail, fail, etc, etc for millions and millions of people.

And somewhere, among all of those millions were two people, so very close together in the merit of their lives. Oh, so very close! Yet their destiny so very different. For one, just scraping through, the joys of heaven for ever. And for the other, eternal damnation.

But it didn't end there. For as thousands and thousands of years pass, and more and more millions queue at the gate, even between these two he must make finer and finer discriminations.

I didn't doubt he could do it, mind you. Well, it'd be more accurate to say that I considered that if anyone could, he could.

But I wondered why he'd want to!

Fifty years on, these are still the two fundamental questions I have about the notion of a standard : the people who define a standard do in fact have St Peter's god-like omnipotence, but do they have his infallibility? And why do they want to engage in a process that is so manifestly unjust?

Order and standard

Let's go back a bit and tease out this relation between standard and rank order of merit. A relation that I intuited at fourteen, but only recently have systematically thought through.

The relationship is not immediately apparent. There are some judges who are adamant that they can recognise standards and this has nothing to do with relative merit. In fact, to them the word relative is anathema. For them, standards are absolute. They are as solid as a winning post, they are a fact established, a sign as recognisable (to them) as a green light at an intersection. Recognising that some people play games, run races, create rank orders and random distributions and normal curves, they see themselves doing work of a higher order; as maintaining absolute quality in a world trivialised by concepts of the average, the normal, the relative.

So let's push them with a bit of Socratic dialogue. Or is it Hegelian dialectic?

> *You can recognise the standard?*
>
> *Yes.*
>
> *Could you always recognise it?*
>
> *No.*
>
> *So how did you come to reach this state of clear recognition?*
>
> *Through many years of study, reflection, and discourse with other scholars and experts. The senses become refined, the observation sharpened, the criteria established, as slowly, with increasing precision, the standard for quality becomes defined.*
>
> *Let's assume all this is true, and you can in fact recognise the standard. So if I were to show you a work that was well above the standard, you would recognise it as such?*
>
> *Of course.*
>
> *Similarly, if you were to be presented with a work well below the standard?*
>
> *Naturally.*
>
> *It would, of course, be apparent that the first work was better than the second work.*
>
> *True. But this is a consequence of my recognition of the standard, and has nothing to do with its cause. It is, you might say, an irrelevant corollary.*
>
> *Possibly. Now let's take a work that is very close to the standard. You would know whether it was just above or just below, would you not?*
>
> *Yes, I could make that judgment.*
>
> *And if I were to present you with another work very close, you would know whether that was just above or just below?*
>
> *Certainly.*
>
> *So if one were just above the standard and one were just below, and I were to present*

you with a third work somewhere between these two, you would know whether is was just above or just below the standard, and you would know that it was between the other two in merit?

I would know that, but only by comparing them all to the standard. Not by comparing them to each other.

Quite so. Now we have talked about five pieces of work. So if I were to present these five pieces of work to you again, you would of course give the same decision regarding each of them.

Certainly.

And incidentally, after the event in your view, you would have them in the same rank order of merit.

Agreed.

Now if they were in a different order of merit the second time, would this not show that there was no absolute standard to which you were able to compare the works?

It would certainly throw doubt on that contention.

And if you can do it with five, in principle you should be able to do it with fifty?

If necessary.

Or even five hundred or five thousand?

Some public examiners do indeed take on that sort of responsibility.

Can we agree then, that regardless of whether the rank order of merit of the works is produced after they have been compared to the standard, or whether the standard is constructed as an artefact of the rank order of merit, in either case the whole notion of standard is in jeopardy unless the rank order of merit is a stable one.

This would seem to be a valid argument.

Would you be willing to put it to the test then?

Put what to the test?

Would you be willing to rank fifty pieces of work in their order of merit, (based on their respective distances from the absolute standard) and then do the same task six months later.

Me personally?

You personally.

I'm a very busy person, and it would quite frankly be a waste of time. The result would *be obvious. It is self-evident. The orders of merit would be the same.*

You're certain of that?

As certain as I am of my professional competence.

Now it is apparent that this whole dialogue is in the Judge's frame of reference, and in that frame the notion of an absolute standard logically implies the notion of a stable rank order of merit of all work samples compared to the standard.

It is also clear that the last sentence is not just a rhetorical device, an appropriate metaphor. It is rather a literal truth specified by the very role of Judge. The whole notion of professional competence is dependent on this ability to judge the value of work in the area. To question that competence, then, is to remove the very foundations of the Judge's professional existence. It is an act, therefore, of extreme danger that we would expect to be resisted with great strength, and considerable emotion.

Quality or boundary

In practice our confidence in the standard defined by a Judge cannot be greater than the accuracy with which the Judge can place works, performances, or people in a stable rank order of merit. Our confidence can, of course, be much less than that, but it cannot logically be greater.

That being so, we may think of the standard in two ways: as the lower limit of adequacy, or excellence; or as the line that divides, as the boundary between classifications. Which way we see it is more than a trivial semantic difference. It is an essential point of discrimination between the frames of reference of the Judge and the General, which entail quite different conceptions of the task being undertaken.

For the Judge claims to judge quality, and if necessary the classifications of quality (as inadequate, or good, or outstanding), and the stable orders of merit are a consequence of this.

In the General frame these claims of the Judge are denied. In this frame it is assumed, and the assumption has much empirical evidence to support it, that a judge produces different rank orders of the same works at different times. This indicates at the least considerable fuzziness of standard, and at the most a disintegration of the very concept of the standard. In addition, different judges produce very different rank orders, as well as very different "standards" around which they appear to be, rather randomly and quite widely, distributed. So in the General frame the first task is to stabilise the rank order as much as possible, and then decide the cut-off, the boundary between the classifications of adequate/ inadequate or whatever.

The point that I want to make here is that these two frames of reference are not compatible, and cannot both be used in the same mechanism of assigning a standard without introducing an inherent contradiction into the whole process. The frames are of different logical types; the Judge is a member of the General class. So contradiction is inevitable when the discourse boundaries between them are not clearly separated.

More specifically, we cannot use the General frame of reference to obtain a more stable rank order of merit, and then use the Judges frame of reference to decide the standard, by looking, for example, at some examination papers around what is assumed (from the General frame) to be close to the boundary line. For the use of the General frame has assumed that any judge is inaccurate, and has already produced not a boundary line, but a broad boundary band, within which the Judges' (many and varied and implicit) definitions of standard are to be found.

The price we have paid for the more stable rank order is to make clear the instability and variability of the Judge's "standard." We cannot now go back to the Judge to determine the many (disguised as the few) indeterminate cases by using his/her ability to recognise the absolute standard, an ability already discredited by the assumptions used to make the rank order more stable.

This has not deterred public examining authorities and professional test agencies from doing just that.

Empirical evidence

Facts are less dangerous than theory; despite the promise of the Enlightenment, most people use up far more energy defending their mythologies than in searching for facts; the world is full of answers looking for questions, and significant questions are rather an endangered species.

There is no doubt about the empirical evidence available about the extreme vulnerability of any single Judge in determining either a stable rank order in concurrent rank orderings of the same tests, or in the great differences in rank orderings between different Judges. And this is just for marking. (Hartog, 1936; Cox, 1965; Rechter, 1968; Halpin, 1983)

On the other hand, those plain statements are sanitised by such mathematical constructs as reliability coefficients, some of which become acceptable because they are higher than others; certainly not because they have solved the problem of the stable rank order. In the literature, reliability coefficients of 0.7, and validity correlations of 0.4, are considered very good. They don't look so good when we realise that 0.7 is fifty percent better than chance, and 0.4 is only sixteen percent better than chance.

Now I want to focus on just one aspect of this issue, which relates to the increased stabilisation of rank order obtained through standardised marking procedures, and show how such collusion of Judges produces confusion in the

General frame.

The fool-proof marking scheme

The Judge's sense of infallibility in his own ability to recognise standards does not extend to his view of other Judges. It can't, of course, because some of them will disagree with him and then they can't both be infallible. It is necessary then in any particular situation for one Judge to be infallible for all other Judges to be fallible. Thus the requirement in any large scale marking exercise to have fool-proof marking schemes, devised, or at least accepted, by the chief Judge.

In this way the lesser Judges take on some of the aura of perfection of the Chief Judge. And certainly, such schemes do have a considerable effect in stabilising the rank order of students being assessed. And of course, it is easier to determine the detail of such marking schemes in such subjects as Mathematics and Physics than it is in English Expression and Art and History. At least one unused to the cognitive gymnastics of examiners might tend to so believe.

Regardless, a Chief Judge who sets a test paper and then devises a marking scheme could, one would hope, be fairly specific about what content and processes were important, and what criteria were being used to assess the students. These particular values, or prejudices, or idiosyncrasies are then passed on to the other Judges through the marking scheme.

It is obvious that this will decrease the differences between rank orders when papers are marked by different lesser Judges. Statistical data can then be produced showing how "good" marker reliability is. And within the Judges frame it is certainly true that rank order discrepancies have been reduced.

What is not so immediately obvious is that within the General frame the discrepancies have been increased. Within the General frame the rank order shows less variation the more independent Judges there are. The whole point of having many Judges is to "iron out," to balance out, individual discrepancies and prejudices. By effectively reducing the number of independent judges through the marking scheme, the generalizability of the rank order produced to another similar situation is reduced, not increased. For example, we can easily imagine another Chief Judge, with different priorities about the course of study being tested, and different criteria for assessment, producing a very different marking scheme, which would then produce a quite different (though equally consistent) rank order of students.

This problem is not solved, though it may be slightly alleviated, through a more "democratic" production of the marking scheme under the eagle eye of the Chief Judge. The hierarchical structure of the committee, the press to conformity and the expectation of a consensus, will necessarily erode genuine independence on the part of the lesser Judges. Regardless, such "consensus" is not equivalent to the averaging out of independent judgments.

Quantum of error

The Judge can be very specific, at least rhetorically, about what is being assessed. And then the error, as defined by the differences between the rank order produced and that of other independent Judges, is large.

In the General frame, we can reduce the discrepancy between rank orders by averaging out the rank orders produced by a number of independent Judges. But then, because they are individually emphasising different criteria, we cannot be very specific about what we are measuring.

Test agencies and Public Examination systems always assume they are measuring what they are being paid to measure, so regard any improvement in stabilisation of the rank order as a good thing. Persig (1976), in Zen and the Art of Motorcycle Maintenance, assumed that this more "stable" rank produced by averaging was indeed a measure of the elusive "quality" which he sought. I find such interpretations exceedingly suspect, examples of wishful thinking.

The fact is that the more precisely we proscribe one aspect of the intricate web in which the spider variously called achievement or ability or quality of performance lies hidden, the more diffuse other aspects become. We tighten up marking schemes and lose generalizability to other marking schemes. We use many judges and lose specificity about what it is we are measuring. We specify behavioural objectives and lose definition of problem solving. We use multiple choice answers and construction and synthesis gets lost.

We create a test and lose most of what we are trying to test.

This sort of phenomena is well known in the sub-atomic world. According to Heisenberg's Uncertainty Principle, you can know the exact position of a particle, but then you lose information about its momentum. Or you can know its momentum, but then lose information about its position. And the amount of fuzziness, the quantum of error, is a constant. A reason for this is that to collect information about sub-atomic particles, they must be interacted with in some way. And the very process of interaction produces a change in the "original" state.

We are in an analogous situation with tests. The very process of giving a test displaces the person from the "original" situation that the test is meant to describe. We have created an interference by the very process of the experiment, and in so doing have activated an irreducible quantum of doubt concerning our "measures," that can never be appreciated by examining just one measure. On the contrary, reducing the error in just one measure may necessarily increase it in another area. For example, reducing the error in rank order may necessarily increase the error in sampling from all aspects of achievement.

Probably the biggest contribution to this quantum of error is to be found in the boundaries of the test situation itself, regardless of the frame in which it occurs.

Such boundaries represent a separation from the everyday learning or working world in which people interact in particular contexts. Knowledge is not something a person has, but rather one aspect of a response, appropriate or not, to a particular environmental context. Test situations invariably remove the person from that real context to produce some sort of controlled, simulated, and hence different context. It is this largely unexamined and unestimated discrepancy that represents a large and irreducible portion in the quantum of doubt.

The enormous popularity (as distinct from reason or purpose) of tests is to be found in its point of congruence with most other myths; in its implicit promise of deliverance from a world permeated with uncertainty, in it's claim to reduce human complexity to a simple story line. In this case the story line of simple numbers.

Judge and jury

> *You haven't really discredited the Judge, you know.*
>
> *I haven't?*
>
> *Of course you haven't. All you've done is to show that some judges aren't as good as they thought they were, and that anyone can be a judge so long as they know something about the topic they're judging on.*
>
> *So I haven't really got rid of the Judge?*
>
> *Not really. You've just democratised the process of judging. You've let more people into the club, and then asserted that the average of their marks is a better estimate of the true score than the judgment of any one of them.*
>
> *You think I've become a victim of my own ideology?*
>
> *Let me put it this way. If you're convicted of murder, does it matter whether the Judge or the jury convicted you?*

Maybe the metaphor is appropriate. After all, the jury has to make a decision. That is its structural obligation, its very reason for existence. Guilty or not guilty. Those are the choices. So someone, either the Judge or the jury, has to draw the line. After all, they either did it or they didn't. There is a truth to be found. And the Judge or jury's task is to find that truth. Who said that?

The error and the standard

It's at this point that the metaphor becomes shaky. For whilst there was indeed a real crime in the case of the criminal, as evidenced by the dead body of the victim, there is less evidence that there is a real order of merit, a true score. Now if there isn't a true score, then necessarily there can't be a true standard. And even if there is a true score, it doesn't follow that there is necessarily a true standard. As we have seen, the error in the estimate of the standard can't be less than the error is the estimate of the true score. And it will certainly be more, because different judges will differ about where to put it.

> *Ok. So why don't we reduce the error in the standard the same way that we reduced the error in the rank order?*
>
> *How would we do that?*
>
> *Get a number of judges to identify the standard, and then average them out.*
>
> *You mean assume there is a true standard, and then see how well we can estimate it?*
>
> *Isn't that what we did with the rank order?*
>
> *Certainly.*
>
> *Then why not do the same thing with identifying the standard?*

Now this dialogue worried me a bit when I first wrote it, and it took me a while to ferret out what was wrong with the logic.

Let's start from the beginning. In the General frame of reference, we assume there is a true score, which mirrors a true attainment, or ability, or trait, or predisposition, or whatever And starting from that assumption, we can show, both theoretically and empirically, that we can never measure it. We cannot specify what it is. We can never specify the true rank order of merit. We can only obtain estimates of it, and indicate how far away from our true rank order it probably is.

Now whether there is "really" a true score or a true order of merit of the group being assessed, must forever remain moot. Assumptions of theories do not have to accord with some relationship between variables that have substantive existence in the world. So assumptions of theories related to people do not necessarily relate to any actual qualities or measurable quantities or substantive aspect or observable behaviour of real people. Theories are useful or not according to whether their outcomes, their conclusions, have some links with the observable world. Their assumptions are just that. Assumptions.

However, if we had clear evidence that the assumption was incorrect, then there would seem to be an inbuilt contradiction of our theory to the world that it purports to mirror.

Now if we wish to use the General frame of reference to define the standard, we need to assume that the rank order is the true rank order. For the true standard requires that preliminary assumption.

The claim of the Standard is not the claim of a broad fuzzy space, but of a thin red line. The Standard, if it means anything, means a point on a stable steel scale, not a probability on shifting beach sand.

Defining standards

And we have seen that we can never present the judge or jury with that true rank order. Our own theory had negated the possibility of locating the standard, because it has negated the possibility of finding the true rank order of merit on which the delineation of the standard, in this frame of reference, depends. It is not moot whether the true order of merit had empirical existence. It does not.

> *Well then, it looks as though we're stuck, doesn't it?*
>
> *What do you mean, stuck?*
>
> *We can't use our rank order, inaccurate as it is, to find a standard.*
>
> *Not altogether true. We can define the standard in terms of our true score. In terms of our true rank order.*
>
> *Whose existence is still moot.*
>
> *Exactly.*
>
> *How do we do that?*
>
> *Very simply. If we wish to use grades, for example, we can just define an A as any score or rank order in the first five percent, and an E as the bottom twenty percent, of the population we are testing.*
>
> *Why five and twenty?*
>
> *Make it twenty and five if you like. It doesn't matter. It's arbitrary. The important thing is to define it, so that everyone is talking about the same thing when they're talking about the grade.*
>
> *Won't there be an error in the definition?*
>
> *Not in the definition. The definition is in terms of the true score. So it is exact, as a Standard must be. Of course, in practice there is always an error.*
>
> *So each person is truly at some Standard, but we can never be sure exactly what that Standard is?*
>
> *The second part of your sentence is true. The first part may be true, or false, or just a silly question.*

Reducing absurdity

Let's briefly summarise what we know about standards, and their relationship with assessment, to this point. First of all, we know that empirically an individual judge cannot consistently recognise a standard, nor can he consistently rank students in the same order. These differences between rank orders, and the position of the standard related to them, are increased if different judges are asked to recognise a standard, or rank order students.

The claim of the Judge that he can do these things is thus seen to be untrue as an empirical fact in the real world. It is a fantasy that he has about his own ability that is shared by many people in society. This does not make it less untrue. It does make it less likely that he will admit to its untruth, and more likely that he will take strong measures to disguise the extent of its untruth. For to admit of any error is to destroy the fragile fabric with which the myth of his power and perfection is woven.

In the General frame the error is admitted, though the assumption of an (unattainable) true score is retained. The estimate of the true score is improved by averaging scores from a number of judges. This is vindicated empirically because different estimates obtained by this method are closer together than estimates made by two single judges.

In this frame, it is admitted both theoretically and empirically that any rank order of students is not the true rank order, but an estimated one with built-in error. Thus it makes rational sense to define some standards, some grades, which admit of no error, in terms of percentiles of this true rank order. Even so, in practice we would have to indicate clearly the errors in our estimated grades. And we would have to indicate clearly that these standards are unrelated to any judgments of "quality" as defined by Judges. They are merely cut-off points at various percentiles of a specified population of testees.

What would not be rational would be to get judges to estimate the cutoff points for standards by presenting them with a scale that was admitted to be inaccurate. The Judge claims to recognise the standard, and the production of a stable rank order is a necessary corollary of that claim. We have rejected that claim in our production of a more stable, but still inaccurate, rank order through gereralizability assumptions. It is absurd to now reinstate the judge to determine the standard. It's asking the judge to do something that's demonstrably crazy.

(Not that it's unusual to engage in crazy activities. It would surely be utterly irrational to expect humans to act rationally. The expectation of rationality is the epitome of delusion. It can lead only to despair at the human condition. To applaud rational behaviour in its rare moments of emergence from the mire of human craziness will provide a firmer path to human happiness. But that's another story.)

Judgments and categorisations in the qualitative world

One more point needs to be made here. Whilst the above argument has focussed on tests and grades as a particular sort of educational event, the arguments made are equally cogent for all categorisations of people, whether these be made in the numerical world of quantitative assessment, or in the more linguistic world of qualitative assessment.

Let us be clear about this. If at any point a qualitative assessment engages in a categorisation, a separation of two groups of people, then it is invoking the notion of a standard, and of the measurement of that standard. And in so doing it is logically engaged in all of the rank ordering and judgment errors that have been discussed.

There are some few genuinely dichotomous variables on the basis of which most people may be categorised; for example, blue eyed people and brown eyed people. Most variables used for categorising people however are continuous and not dichotomous; as such, any such categorisation requires a standard, the thin red line that defines the categories, and then a judgment about whether any particular case is above or below that line. As argued earlier, this logically implies a stable rank ordering, which constitutes a primitive form of measurement. Categorisations then involve both standards and measurements, regardless of how much semantic camouflage is used to disguise this.

Democracy and doubt

As the judge topples from his autocratic pedestal of certainty, it is doubtless pleasing to those of democratic mind to know that what will replace the judge is not chaos, but the will of the people. The rule of the individual will be replaced by the judgment of the group. The idiosyncrasy of the individual will be cancelled out and reveal the pure decision of the majority that is the source of the true the right and the just!

We have seen how in practice the delineation of the standard cannot be more specific than the fuzziness of the rank order of those being standardized. And we have seen how individual judges vary considerably in their rank ordering of a group of students, especially if they have no information about them other than the set of examination or test papers. A good punter can (usually) pick a good horse from a bad one, in a general sort of way, but he makes lots of errors when trying to rank accurately all of the runners in a particular race. So it is with the judge of human performance.

There is a crucial difference between the punter and our Judge, however. In the horse race the camera can photograph the finish, so that there is a "true" rank order in which the horses run this particular race. It might not be stable if they run this distance next week, or generalizable to other distances. It will certainly be different over hurdles. But at least in this race we know accurately what the rank order is. Further, we know (almost) exactly what distance they have run, because we have a unit of distance with which we can measure. And we know (almost) exactly what time each horse took to run this distance. If we wanted to, we could nominate a "standard" for this distance below which horses could not compete in the equestrian Olympics. It would be an accurate standard. And it would be arbitrary. And we could measure whether a horse had reached that standard with a small, and empirically determinable, error.

Horse racing as we know it is not a good metaphor for the testing game. So let's develop a better one, a myth more appropriate than that of the infinitely accurate little black box that had mystical knowledge of standards, and resides in the head of the omnipotent judge.

They're racing in Testland

In Testland, races have always been important events. There are no permanent tracks, and unfortunately no way of measuring either distance or time with any accuracy. Some of the more exalted people in Testland do own clocks, but unfortunately they all run at irregular rates, and they all give different times for the same race.

Races are accompanied by due pomp and pageantry. The track is marked with flags and signs saying "this way" and "that way." Horses and riders train hard and are decorated in much colourful finery. There is no starting point and no finishing point but when the bugle sounds they are off and may the best horse and rider win.

There is no actual finishing point, but everyone knows the general area that the race will finish. Here congregate the Judges: the Standard Judges in their white wigs and purple cloaks impressively flourishing their clocks; and the Placement Judges so serious in their blue serge working suits all constructing their own lines of sight so they can accurately record the order of finishing. Some of these, aware of the subjectivity of human vision, have cameras with which to record the finish in a truly objective way.

In the good old days in Testland there were many more Judges than horses. Everyone would have a great time picking the winner, and recording the orders and times. Then they would happily argue for the rest of the day about who had won and come second and so on. Because all of the judges were viewing the race from different positions and at different angles, because it was unclear which part of the horse had to get past the finishing line to complete the race, and because the signs on the track often had horses running in opposite directions by the time they reached the finishing area, every rider could find some judges who thought they had won the race. So race days were days of celebration and festivity, until . . .

Nobody knows quite when the rot started, when the question about who really won the race became a problem for decision rather than an excuse for argument. Some thought it was when someone suggested that prizes should be given only to the first three horses and not shared equally as was the custom. Others thought it stemmed from a misunderstanding of a remark made by one Sir Henry du Princely, the Queen's sometime lover; another Judge thought Sir Henry said he had the best clock in Wonderland, and took umbrage. But most saw it as the inevitable march of progress and civilisation as Testland lurched forward into an uncertain future; just another example of the dominance of the three e's in the post-industrial era; engineering, efficiency, and expediency.

Regardless of the reason, the facts are clear. Word got around that there was a real winner, and a true rank order in the race. There had to be, because it was self evident that some things were better than others. It followed that some horses and riders were better than others. Thus no-one but an idiot would argue with the blinding clarity of the truth that there was a unique winner, and a verifiable placement order, to every race. The race, everyone knew, was to the swiftest. It became the task of the Judge, therefore, to determine that swiftest.

Sir Henry, who had the ear, as it were, of the Queen, and had been under some flack from other Judges because of the misunderstanding previously referred to, made a unilateral decision that henceforth and from hereon only one clock would be used in adjudging horse races and that one would be his. One or two other Standards Judges who contested this pronouncement found that their clocks mysteriously disappeared, leaving them, clearly, without a tick to tock on, or alternatively a tock to tick on, depending on which University in Testland you went to.

Changes of this magnitude are not implemented easily, of course. At the next race meeting Sir Henry clocked the winning horse and for obvious reasons no other Judge queried his timing. However, the Placement Judges argued that, through no fault of his, he had clocked the wrong horse. Obviously, Sir Henry had underestimated the complexity of the task. He needed the placement Judges in his pocket as well as his clock.

It was at this point that Sir Henry's brilliance shone through with a remarkable insight which ensured his historical survival in the annals of Testland. He let go a double-bunger of a pronouncement that in one foul swoop solved the otherwise irresolvable time and space problems. He defined the finishing line as being where his clock was, and in the direction in which he pointed. By these means Sir Henry succeeded in defining a unique standard and producing a unique placement system at the same time. Truth was now defined. It was what Sir Henry did. He had constructed a new view of reality. A world of winners and losers, scientifically classified.

In conclusion

The astute reader will recognise here the birth of the Judge's frame in its modern form. More importantly, they will see, from their helicopter oversight, that the race has not changed. From above the chaotic nature of the race is evident, and Sir Henry and his little team of supporters can be seen to be doing what they are in fact doing; co-creating a fantasy about a winner where there is none, blinkering vision to substantiate a myth of order, and imposing truth by political assertion.

Chapter 12: An Inquiry into Quality

Synopsis

From the last two chapters it becomes evident that a fundamental purpose of relating assessment descriptions to standards is to transform notions of quality to notions of quantity. So in this chapter the notion of quality is discussed, and some of the differences with the notion of standard are elucidated.

The theory of logical types is briefly explained in terms of its implications for complex constructs with multidimentional aspects and the special properties of the class "safety standards" is discussed.

The construction of a bridge with various criteria for quality is discussed to illustrate the different languages that must be used to justify the quality characteristics for each criteria. The subsequent history of the bridge is then used to illustrate how the notion of quality is related to boundary conditions and events, and how this affects notions of permanency and attribution.

Some reflections on the nature of quality follow. These are then applied to some of Eisner's ideas about connoisseurship.

Persig's ideas about the metaphysics of quality are briefly discussed, and the relationship between morality and quality on the one hand, and static and dynamic morality, introduced.

All standards are arbitrary

When I was younger and groping for a profession that might suit me, I studied Physics and Engineering. I don't remember much of the detail of those studies, but I did learn two things that are pertinent to this chapter: One is that all measurements contain an error; the other is that all standards are arbitrary.

I remember very clearly struggling with some calculations to determine the cross-sectional area of a steel beam for a bridge. Estimations of maximum loading on the bridge, moments of force and tensile stress resulted in a value of the cross sectional area of the beam accurate to three figures. However, before choosing the appropriate steel T section there was one more step. A safety factor of three must be applied. Or was it four? No matter, the calculated cross-sectional area must be multiplied by this arbitrary number in consideration of possible tornadoes, earthquakes, rock concerts on the bridge, or whatever other natural disasters might inadvertently occur. This undoubtedly would make the bridge safer for traffic and incidentally more profitable for the steel manufacturers. And it made the accuracy of the initial calculation absurd.

Safety and quality

At this point I want to try and untangle another confusion that has bedevilled the notion of standard, especially as applied in the human sciences. This is the confusion between safety standards and quality standards.

In the manufacturing area there is less confusion. Standards that apply to car seat belts, bumper bars, brakes, lights, are clearly basic safety requirements. General design of car, colour, control panel layout, type of upholstery, fuel economy, are aspects of quality. And of course, one aspect of quality is that all safety standards are met.

Safety is about prevention. Safety is about what is not, about events that are always immanent, yet, if safety is successful, never materialise. Safety is about the future that is frustrated, about unrealised potential. Because each safety measure blocks a road to disaster, each safety measure is essential in its own right. To meet a safety standard is to claim that one such roadblock is in place. To know that all such safety standards are met is to be reassured and insured against disaster. However, to know that eighty percent of safety standards are met is to know nothing about which particular safety standards are not met. For a gambling man this may be a situation of high desirability, and hence provide an experience of high quality. But in the world of safety standards, this is a recipe for disaster.

Quality on the other hand is about manifestation, about potential realised. Quality is not so much about specific aspects as about their interrelations; about interpretation rather than measurement; about the whole gestalt rather than summaries. Further, notions of quality are intimately and necessarily connected with the observer, and hence are constructed from the observer-object interaction, rather than claiming to be a measurable component, or sometimes a presence or absence, of the object or specific attribute being observed.

Theory of logical types

The theory of logical types is about levels of abstraction in human discourse. One of its axioms is that whatever involves all of a collection must not be one of the collection; that is, that there is a fundamental distinction between a class, and the members of that class. This might seem obvious. Obviously a single man is not all men, and a married woman is not all women.

Trivial as this might seem, the conclusion from the theory is far from trivial: that when this clear separation between class and members is not made, messages become confused. As Bateson (1972) describes it, "the theory asserts that if these simple rules of formal discourse are contravened, paradox will be generated and the discourse vitiated" (p280).

Human discourse is decidedly more complex that simple logical syllogisms. We do not usually talk like logic machines. We talk very often in and about abstractions, and these abstractions may be at different levels of logical type. We present information (first level), and give an interpretation of that information (second level), in a particular context which affects its meaning (third level). A story that makes fun of a rich Jew has a very different meaning if told by a speaker at an anti-semitic rally than it does when told by a Jewish comedian on a New York stage.

Of particular interest here is that errors that lead to confusion occur when the properties of a class are ascribed to members of that class, or vice versa; or more subtly, whenever the discontinuity between class and member is neglected, and they are treated as if they were at the same level of abstraction:

> The theory of Logical Types makes it clear that we must not talk about the class in the language appropriate for its members. This would be an error in logical typing and would lead to the very perplexing impasses of logical paradox. Such errors of typing can occur in two ways: either by incorrectly ascribing a particular property to the class instead of to its member (or vice versa), or by neglecting the paramount distinction between class and member and by treating the two as if they were of the same level of abstraction (Watzlawich, 1974, p27).

Safety and logical type

Safety is not quality. It is one criteria we might use in describing quality. It is a member of the class of such criteria. But it is a very particular member, because it is atomic in its construction. It is comprised of a number of specific safety requirements each of which must be individually met. Not only is the class of events or information called "safety" of a different logical type to the class called "quality," but the essential information about safety is lost when the class "safety" is described, rather than the individual items that describe it. Unless, as we mentioned earlier, the statement about the class is that "all safety measures have been satisfied."

Safety and people

In many aspects of our life safety measures are important for its continuance. In home, leisure activities and job, safety requirements contribute to our health and that of others. So matters of safety are a part of various educational programs. As such, it would seem important that evidence be obtained that students have incorporated such safety items into their behaviour. Or, at the very least, that they understand and can implement all of the safety requirements. Talk of safety (like talk of sexuality) produces points of high density in the field of power relations.

It should be apparent, however, that test or examination information involving rank orders or grades or marks regarding safety represents information about the class of safety items, and as such is inappropriate and confusing. If safety requirements are essential requirements, then marks of 70 per cent or grades of C for safety, or for tests which include questions about safety, present information that is inherently contradictory. By definition, if you have not met all safety requirements you are unsafe.

Test-makers and others argue that in the context of a test people make errors and it is not reasonable, because it rarely happens, to expect one hundred percent correct response. This is surely an indication that the test context is inappropriate for obtaining information about a person's acquisition of safety measures. It certainly does not justify accepting that if they can provide evidence that they "know" seventy percent of the safety requirements that their "standard" of safety is adequate.

Further, talking about safety measures, or choosing the correct safety requirement from a number of choices, is an activity of different logical type than implementing that information in the context of a job. Talking about something you do is of a different logical type than doing it. So any measure on a test, even at one hundred percent, cannot be a measure of safety behaviour. It is a measure of test behaviour. At the very best it is an indicator, about which empirical evidence could be obtained about the probability of its correspondence with overt safety behaviour under specified conditions. In this respect, probabilities less than one would necessarily indicate test invalidity.

Safety and minimal outcomes

The idea of minimal outcomes is analogous to that of safety. Minimal, or minimum, means the least amount, the lowest possible. If a course of study has a set of minimal outcomes that define its successful completion, then by definition all such outcomes must be demonstrated if the course is to be satisfactorily completed. To set a test incorporating questions related to such outcomes and then use a test score (a statement about the class) to describe the "standard" that has already been described by each of the members of the class, is again to confuse logical types. Such tests are sometimes referred to as mastery tests.

There are three additional confusions, two of them the same as for "safety." The first is that only a perfect score is consistent with the definition of minimal. So to attempt to find an appropriate "cut-off" score to use as a standard is to engage in a paradox, is to indulge a contradiction, is to professionalise an absurdity. Berk (1986) was able to identify 38 methods for setting standards and produced a consumer's guide (to choose the most appropriate absurdity).

The second confusion involves the fact that context affects meaning. For many educational outcomes the context of a test situation is inappropriate anyway and represents another logical type confusion. For example, any outcomes involving verbal discourse, such as listening skills, group problem solving, giving instructions, cannot be demonstrated in a written or multiple-choice test without logical type confusion occurring. Writing about verbal interaction is not verbal interaction. Choosing the most appropriate response from a multiple-choice selection is not responding oneself in an interpersonal context. Talking about a painting is not painting. The whole test and examination industry is permeated with this sort of confusion.

The third confusion is one of ends and means, and is well described by Burton (1978): "no measure of a single skill can ever be mapped on a non-trivial vision of real success because any problem can be solved in more than one way. One can determine whether the respondent has the skills necessary to solve the problem this way, but one lacks the justification for imposing successful performance, this way, as a standard"(p273). Burton believes that "this argument is fatal to any method of setting performance standards." Burton is perhaps mistaken in believing the issue is amenable to rational argument, and does not consider that it may be entrenched in mythical discourse.

Mastery tests and frames

Mastery tests result in scores produced by the summation into a numerical score of specific objectives attained. In relation to error, they contain all of the errors of specific objectives plus a large labelling error. In adding the results most of the important information is lost, in that we no longer know which specific objectives have been attained and which have not.

In this situation, whilst the generation of the test has used the Specific frame of reference, the summation has resulted in a normative test score. We no longer have information about what a student has achieved. We have information only about how many of the objectives have been achieved. This is exactly equivalent to information about how many addition sums are correct, or how many words are correctly spelt, or how many formulas in dynamics we can remember. The description is now clearly normative, and may only be interpreted in terms of whether one student got more or less "right" than another, or in terms of some arbitrary "standard" of how many "correct" answers will be considered "adequate"; how many correct answers constitutes a "pass."

In this situation, because information about the particularity of objectives attained is lost, the whole detailed descriptions tend to be similarly "lost," or unavailable to those interpreting the test information. Labelling errors thus become large, as the meaning of the score, and the label attached to it, are differentially interpreted.

Mastery tests and internal logic

In most courses there are some facts, some understandings, some activities or skills, which are central to what the course is about, so that we could say - if they don't know at least those things, or if they can't do at least these things, then there is no way we could say they have adequately completed the course. In old-fashioned terms, they are the "must knows" or "must dos" of the course. As distinct from the "should know" or "could know" categories.

Now there may be some areas of study where curriculum writers or teachers are unable, or unwilling, to specify such a category of "must know" performance. However, when it is so specified, it comprises a description of a finite number of procedures or products that will demonstrate the "knowing" of these crucial things. In other words, within this limited "must know" area, it is possible to specify what must be done, the conditions under which it must be done, and the procedure by which its adequacy will be known.

These then could be used to describe the essential requirements of the course of study. They are limited in number and extent, and are specifiable in the specific frame of reference. As they are accomplished, as evidence is obtained that each outcome has been achieved, this can be certified by the teacher or student. If there are ten such outcomes, then successful completion of the course would require that all ten outcomes be so certified. Otherwise they cannot, obviously, be essential. To certify that eight out of the ten essential requirements have been completed is to certify that two of the essential requirements of the course have not been completed, and thus to certify that the student is uncertifiable. More than this, it is to lose the information about which two essential requirements have not been demonstrated.

So to obtain a "total score" on a mastery "test" is to contradict the whole concept of essential requirements, and to lose all the relevant information. Unless the total score is a "perfect" score.

In many situations the very notion of a "test," of some particular situation constructed to check all of the essential requirements at one time, would itself be contrary to this frame of reference. In the artificial and often pressured "test" situation it might be expected that success in some "essential" activities might not be demonstrated. It is this very argument which has been used to justify the acceptance of less than a "perfect" score in a mastery test. Rather it should be seen for what it is - an argument that invalidates the use of the test.

The problem of time-binding is not solved by success in test situations any more that it is by success in the ongoing teaching - learning context. We can never certify that any fact will be recalled at a later date, that any understanding will be retained in the future, that any skill will be demonstrated again successfully next year. We can sensibly certify that a behaviour has occurred once, or twice, or if necessary one hundred times. Regardless, we can never be certain it will be adequately demonstrated on the next occasion.

Test givers imply, with their insistence on testing, that demonstrations outside the testing situation are in some way of limited value, credibility and validity. It has always seemed to me that "tests" have all the inadequacies of "on site" or ongoing certification, with quite a few bonus inadequacies added on for good measure.

Or more accurately, for worse measure.

A bridge of quality

Let's assume that we want to describe a particular person's performance in a certain area. Building bridges is as good an area as any. And we are interested in the quality of that performance. That is, we are in the area of discourse often called assessment.

We might decide that there are four aspects of performance which we want information about; four members of the class we will call quality; four criteria on the basis of which we will assess quality of the bridge produced. Is the bridge safe? Is it economical in cost of materials, construction, and maintenance? What is its environmental impact in its rural context? And how is its aesthetic design judged in a competitive order of merit in relation to other submitted designs?

We note in passing that this decision about these particular four aspects of quality is itself a value judgment subject to enormous error in the General frame of reference.

It is clear that the language of discourse for each of these four criteria will be different, and attempts to simplify by means of some language that is appropriate to some and not others, or that is appropriate to the notion of "quality" as a class but not to some or all of the members of that class, is to compound confusion by oversimplification (Eisner, 1991, p182).

For example, the first question, about safety, may only be addressed by showing that all safety measures are in place; the language that designates individual safety standards is appropriate. The question about being economical involves careful costing; the language of accounting is appropriate, and the language of economics will be necessary to delineate boundaries. The question about environmental impact will draw information from a number of disciplines - geology, biology, ecology, geography, ethics, economics, and so on. Ultimately, the discourse must deal with the balances and trade-offs among conflicting values and pressures; the language of politics and the language of environmental ethics will fight it out. Finally, the order of merit based on the aesthetics of the design will draw on the language of art and architecture, and be involved with issues of the assessors' personal tastes and the profession's current fashions. Finally, however, such complexities will be reduced to a single dimension where better-worse becomes more-less and a rank order is produced.

As this competitive order of merit is one aspect of the quality of the design, it is not that quality. By the same token, no measure of the order of merit can be the measure of quality, any more than a cut-off point on the order of merit can represent a cut-off point of quality. All this regardless of how consistent, stable, generalisable that order of merit may, or may not, be.

Permanence of quality

> *I've been thinking about the quality of the bridge.*
>
> *The one where I chose four rather arbitrary aspects of quality to talk about?*
>
> *Yeah, that one. You made it easy for yourself by choosing something very practical and material and solid. I mean, it's stable, you can see it and jump on it. It'll still be there tomorrow so that others can assess its quality for themselves.*
>
> *It does have that illusory aspect of permanence.*
>
> *Why illusory? A bridge is a pretty permanent structure.*
>
> *Even so, the notion of quality is somewhat ephemeral. Let's see how our bridge, built five years ago, has stood up to our quality assessment. First the aesthetic quality, the only one subjected to the rigours of competition, of rank ordering and the notion of the standard. The design was brilliant and quite spectacular. There was some controversy after it was built about its enormity. But mostly there was approval. Then, of course, fashions change. Most "experts" these days consider simplicity a major design virtue.*
>
> *You're saying that if the competition were rerun today this design wouldn't have won?*
>
> *That's what I'm saying. These days big high ornate bridges are out. Simple low bridges are in.*
>
> *What about environmental impact?*

> *There's the bridge's visual domination of the landscape, which is much more intrusive than was anticipated. The terrain is very flat. So you can see it twenty kilometres away. But more important for some is the impact it's had on the lesser crested poorigal. The bridge has affected its navigational ability in some mysterious magnetic way. Apparently this area was significant to a change in direction during their yearly migration. Now they fly in circles around the bridge till they drop. Suddenly they've become an endangered species.*
>
> *What about the economic question?*
>
> *Interest rates have gone up by a factor of three, they've put a toll on the bridge, and the government has had to bail out the Roads Board once already. What was once an economic asset has become an money-eating monster.*
>
> *Well, I guess fashion, the environment, and the economy are always a bit suspect in terms of their stability. But at least the bridge is still there, and it's safe.*
>
> *Not exactly.*
>
> *What do you mean, not exactly?*
>
> *Just one of those unfortunate things really. It's not considered a major earthquake area. Almost no activity over the last sixty years. Then last week there was this major tremor. Point eight on the Richter scale. A major fault line developed just a kilometre away from the bridge.*
>
> *Did it damage the bridge?*
>
> *Not exactly. Amazing structure really. Shows how good the design was. Not a crack anywhere. Only one problem.*
>
> *What's that?*
>
> *When the land tilted, the whole bridge tilted with it. The road slopes thirty degrees.*
>
> *So what happens now?*
>
> *Well, the bridge is useless. The only question now is whether to leave it there, or spend half a million to blow it up and remove it, thus saving from extinction the lesser crested poorigal.*

The apocryphal nature of this story does not diminish the fact that the bridge, like everything else which has a material presence on this planet, is not permanent. It will change. It is not fixed in space and time. The rate at which it is ravaged by time - that is, by the events that indicate its interactions with the environment - is normally quite slow, and hence our sense of its relative permanence compared to our own brief life-span. Yet in geological times the life of the bridge, as a bridge, is minuscule.

What is important to understand about this very sad story is that it indicates very clearly that the bridge itself does not have any qualities. Putting it another way, none of the qualities we discussed in relation to the bridge belong to the bridge. They are rather descriptions of how the bridge will interact with other things - with the physical and geological environment, with the economic system utilised to finance it, with the human cultural world in which it is enmeshed. So when any of these environments change from those expected, so does the quality of the bridge.

Nor does the bridge have some aesthetic qualities having a magical existence independent of the bridge and its environment. You may conceive the bridge as being beautiful, as some music that you hear is beautiful, or the second law of thermodynamics seems beautiful. And indeed there may be a palpable human response that you have to these three events which justify using a single word, beauty, to describe them. Even so, it is clear that the similarity is contained in your particular response to the events, rather than to the objects that are responded to.

All of which does not mean that beauty is in the eye of the beholder. To take that view is to denigrate the object. Just as to ascribe the beauty to the object observed is to denigrate the observer. If the label of beauty is to be pinned anywhere, then it must be pinned to the event, the interaction, the relation, between observer and observed. Qualities, like any other form of data, are constructed from events, not discovered in objects.

Quality, standard and logical type error

Let's look then at what might represent quality in a teacher or student in a school.

The function of the school is not only to prohibit and punish and exclude but to produce. To produce good work. Though even here, good work is but a symptom of the more important school product, the good student. The good individual student. Increasingly, it is not so much the work of the student that is valued, but the "whole person" that presages it. Abilities, attitudes, skills, the whole plethora of attributes fantasised to define the good student, the good worker, the good manager, become the focus of attention, the point of application of the standard.

This is not new, though it is more overt that it was twenty years ago. I remember doing some consultancy work in a Primary Teachers College in the 1960s. I visited the various faculties, and talked to the lecturers. Indeed, they were concerned that the students had sufficient knowledge to teach the subject. But what was more important was that they had a very positive attitude to the subject, that they really liked teaching mathematics, or music, or history, or science, or physical education, or whatever. On the surface, a useful intent. Yet when I tried to picture what sort of a person this would be, with great enthusiasms for everything that they taught, I could see a successful sales-person, but hardly a successful teacher.

It was laudable that these lecturers communicate their enthusiasm to their students. It was their inability to see its overall implications, and its curtailment of any critical thinking on the part of the students (or indeed often on their own part) that was cause for concern. My problem was to discern the difference between a student enthusiastic about the whole curriculum, and a happily conforming blob.

The error is a logical type error. In the class "quality" there are many members; there are many aspects of a person that relate to quality performance. One of these may relate to the particular context. Another may relate to standards of proficiency. Another to integrity of values. The language of discourse of these three areas will be different. But all of these discourses must be both utilised and transcended in a discourse on quality, and no measures of the members of the class (assuming such measures are possible), can be a measure of quality.

Another example; quality of life is not the same as standard of living; there is a world of difference, indeed a life-style of difference, in the two concepts. For the very essence of quality is its immeasurability, its identification with a world not wholly material, an association with that mysterious realm of experience called "soul." Quality is concerned both with essence, with experience from within, as well as with experience perceived through reflection from surfaces. Standard of living, on the other hand is a function of measurable quantities; income, savings, washing machines, televisions, supermarket shopping bills, and whatever; the countables, the quantifiables, of the material and materialistic world. Again, "standard" is a member of the class "quality." And for that very reason the two concepts cannot logically, and hence rationally, be identified.

Adequacy and labelling

How do we solve the dilemma? If standards cannot do the job expected of them, what do we replace them with? The issue of competence in a job does not go away because of the errors and confusions in its measurement. On the other hand, it is possible within a particular milieu for a group of people to agree with some consistency, and hence certify, that certain work has been carried out adequately. In every family, in every school, in every sporting team, in every job, work is done and considered adequate. It is useful for some purpose and not dangerous. And the conditions of that work, (and hence of that agreement), may be democratic or elitist, may press towards convergence or divergence. In other words, there is a notion of adequacy, or competence, or comparative excellence - in short, of a limited sort of quality, that is both embedded within and produced by any work culture, in terms of which individual performance is assessed. What is also clear is that this notion is fuzzy and multi-dimensional, error prone, describable rather than measurable.

What becomes clear here is that this notion of adequacy, of quality of the work, is not independent of the culture in which it occurs. The label of adequacy is a label belonging to the whole interactional milieu in which the work occurs; yet another reason for the immense errors that become apparent when such work performances, or the abilities or skills or predispositions or aptitudes that are fantasised to explain them, are pinned onto particular workers, and to a lesser extent on particular criteria or products (Fielding,1988; Raven,1992).

Quality

> Quality . . . you know what it is, yet you don't know what it is. But that's self-contradictory. but some things are better than others, that is, they have more quality. But when you try to say what quality is, apart from the things that have it, it all goes poof! there's nothing to talk about (Persig, 1976).

Maybe the apprehension of quality really is a mystical experience. And maybe not. On the basis of the discussion so far, I will try to give the skeleton a bit more flesh.

Quality refers to a particular experience. The notion of quality is a complex one, involving a number of aspects of the experiential event that can be discriminated. The possible aspects that could be discriminated always exceeds the actual aspects discriminated; an informed choice is made about what particular aspects will be discriminated in this particular case. The choice itself is arbitrary, in that different choices could have been made, some of which would in retrospect be approved. Such choice of course mirrors value.

Discourse about any one aspect might or might not refer to some standard of accuracy or adequacy or competency or whatever.

Balance or harmony or elegance is an aspect of quality. This involves the relationship between the aspects initially discriminated. All this so far is a description of surfaces, of what the object or performance appears to be from the outside.

How does this relational aspect look from the inside? If quality is more the spirit of the product (the person, the event), then quality relates to the interior of the holon. Quality is, in human terms, the expression of the life force immanent in the product, or in the production, or in the person in the process of production; that is, in the production event. Quality then becomes related to a state of consciousness, or its analogue in non-conscious productions. It involves the integrity, the meaning, both of the producer and the product.

Quality also involves the integration of the inside and outside; the aligning of truthfulness with truth; of inside and outside awareness; of the aligning of the potential of the stone with the vision and skill of the sculptor; of the sound of the spirit with the song of the singer (Wilbur, 1996).

From the inside quality is experienced as the essence of the event, of the spirit of the relational experience. It is thus the meaning of the event as interpreted by its participants. It may be, indeed will be, different to other similar eventful experiences, and because of its idiosyncrasies is not comparable to them in any linear way. So it is not possible to link this notion of quality to ideas of adequacy or competence or of other categorisations which necessarily involve standards. What words then are suitable? Beauty perhaps? Elegance? Flow? Life? Spirited? Words that describe the essence of the experience, of the connection!

In relation to people's performances, the notion of quality can be attached either to the creative process of the performance, or to a particular product of the performance. Post-structural analysts want only to attend to the latter, regarding the former as irrelevant. And of course the event that involves a critic interacting with the product is a different event to that event which produced the product. As such the qualities of the two events are necessarily different and essentially non-comparable. The element they have in common is the final product; but this product was the culmination of the first event; it did not exist till the final moment of the first event.

On the other hand, it is sometimes a stable and reproducible element of the second event. The two events are holarchicaly connected. The first event (culminating in the product) can exist without the second (the critique). But the second event cannot happen without the first. It follows, as with all such holarchical connections, that the attributes that determine quality in the first event are not necessarily or probably those which determine quality in the second. They are different creative endeavours; they have different intentions and languages; to misrepresent this difference is to court confusion.

Eisner, quality, judgment and standard

Eisner is one of the few writers in the assessment field who has attempted to analyse in depth the notion of quality through his notion of connoisseurship. Eisner (1991) differentiates qualities from qualitative from quality. "By qualities I mean those features of our environment that can be experienced through any of our senses"(p17). So a quality pertaining to a person is any aspect of that person on the basis of which we can differentiate by using our senses. "Aspect" or "attribute" or "property" may be better words to use because they avoid the confusion with the notion of quality we have been discussing. He goes on to claim that "we can only appraise and interpret what we have been able to experience," but then warns that "if our perceptual experience is aborted for the sake of classification, our experience is attenuated"(p17). Eisner adds that "the qualitative aspects of experience are not only secured in attending to qualities out there, but also are manifest in the things we do and make"(p18). In my terminology, aspects are discriminated both in the event that produces a product, and in the event in which it is perceived.

"The ability to make fine-grained discriminations among complex and subtle qualities" is what Eisner (1991, p63) calls connoisseurship, the art of appreciation. The art of recognising quality, as I am using the term. He recognises a fundamental problem with his notion of connoisseurship:

> we may find critics with very different views of the same situation or the same book. What are we to do with such differences? In standard research methodology, we might dismiss the critics as incompetent and find new ones who can independently agree, or we might look to our own criteria and methods, for these might be at fault. Our methods might not be clear or, if clear, they might be incomplete, or our instructions to our critics (or judges) might be ambiguous. The point is, we would not trust differences of view; such a circumstance indicates statistical unreliability. We would try to achieve reliability among judges. As a last resort, perhaps, we might decide to limit what the critics were to attend to. By simplification we might achieve a higher level of intercritic agreement, even if in the process we compromised validity (p113).

Obviously, Eisner does not agree with this response, and is critical of it. "Critics might be attending to different dimensions of the same work," he points out. They might be bringing different perspectives to it, be sensitive to different aspects of it. No one knowledgeable in literature, "would dream of trying to calculate a mean among critics as an adequate test of a critic's work"(p113). Maybe not, but such consensus is often seen as an adequate test of the work being criticised, and that is the issue here.

And indeed, that is Eisner's test for the adequacy of the critic's work: "consensual validation in criticism is typically a consensus won from readers who are persuaded by what the critic had to say, not by consensus among several critics"(p113). What is such local consensus except a qualitative calculation of the mean? And note how the second order consensus has distracted attention from the first order contradiction, to which he does not return.

Why are collections holding contradictory judgments so difficult for Eisner? In his criticism of specific behavioural objectives, Eisner (1985) says that those who evaluate them "often fail to distinguish between the application of a standard and the making of a judgment" (p115). He then quotes Dewey, who, he says, "makes the distinction quite clear." So what is the distinction according to Dewey? Standards, according to Dewey, define things with respect to quantity. And measuring a quantity is not itself a mode of judgment.

And qualities are qualities of individual objects, even though the critic reveals himself in the criticism. So to Dewey, and Eisner, the qualities are indeed inherent in the individual object, even though the description of those qualities is enlightened by the connoisseur. And nowhere, concludes Dewey, "are comparisons so odious as in fine art" (Eisner, 1985, p115).

So Eisner is clear that qualities cannot by measured by standards. And of course they can't, because standards are definitions and not measurements. What he must mean is that qualities cannot be measured by comparing with standards, both because measurements and judgments are of a different order, and because comparisons are odious.

So he is trapped; qualities are inherent in the object; connoisseurs make the fine discriminations that enable them to describe quality; such judgments are not measurements and abhor standards; even so the judgments might lead to categorisations of the object (of winner of the contest, or worth a distinction, or inadequate at this level), which bypass standards and measurement. Yet connoisseurs differ sometimes fundamentally in their categorisations.

I have argued in the previous chapter that such categorisations necessarily invoke standards, and comparisons with them. But even if they don't, two contrary judgments of connoisseurs create a contradiction that denies that connoisseurs can categorise accurately, and this is surely one of the essential aspects of their connoisseurship. An alternative explanation, of course, is that the qualities do not reside in the object, but are rather an aspect of the event that involves the interaction of the object with the critic. In which case to categorise the object is to mislabel the event, and hence by implication to mislabel the person who produced the object.

All of which takes us back to Eisner's original question: What do we do with such differences? Eisner says don't do what is usually done. And then is silent. Maybe if you ignore them they'll go away! I note that he is talking about consensual validation in this section of the book, and validation, as we have seen, is an advocacy argument for the defence. It follows that the disagreement has to be ignored, because it represents the essence of the (unspeakable) case for the prosecution (See Chapter 16 on Validity).

Summaries or collections - the crucial choice

So Eisner doesn't want to celebrate difference as being at the cutting edge of new knowledge, the collection being the best description, superior not only to a summary, but also to any consensual agreement. For to do this is to deny the possibility of the accurate categorisation of people or their creative products. And that is the cutting edge of the power of the connoisseur. Such power does not ultimately lie in the cogency and plausibility and depth and sensitivity of his critique, however much the connoisseur may wish to believe it is so, and even though this advocacy may well support such power; in practice it lies in judgments that define the standards that produce the categorisations that determine the lives of Jack and Jill and all their little children.

This necessity to categorise in a single dimension is illustrated by Rosenberg (1967). In his book On quality in art, he looks at criteria of excellence from the 16th to the 20th century. He quotes de Piles, a 17th century critic, who:

> evaluates the best-known artists of the past and present in a very special way: the artists are graded in each of four categories already mentioned (composition, drawing, colour, and expression). He

> scores each category against an ultimate grade of 20, which would indicate perfection (p36).

He then goes on to say "de Piles does not give us the sum total for each artist." Presumably it never occurred to him to do so. But then Rosenberg adds: "but we can easily do the addition"(p36). Presumably, as a child of the 20th century, it never occurred to him not to.

Rosenberg (1967) then uses this magical and meaningless sum total to criticise some of de Piles' ratings; "We are disappointed that he rates Michelangelo (37) much lower than Andrea del Sarto (45) . . . We cannot understand why Durer receives a grade of only 36, when a second rate Mannerist like Taddeo Zuccaro gets a total of 46"(p37). And so on. But of course de Piles gave no such grades. He knew it was meaningless to add a mark for colour to a mark for composition to a mark for drawing.

In assessment, whether qualitative or quantitative, the crucial choice made is whether to opt for summaries or summations on the one hand, or for collections on the other: to opt for summaries is to go the way of simplicity, of communality, of "truth." A summary celebrates similarities by defocussing differences; to opt for collections is to stay with complexity, with uniqueness, with essential uncertainty. A collection celebrates differences by defocussing similarities.

Summaries and summations then are basically conservative; they are uni-dimensional; they are dedicated to notions of order and security. Collections are basically radical; they are multi-dimensional; they are dedicated to notions of creativity and anarchy (in its positive persona).

To date, the history of educational assessment has been a developmental history of the summary. The current agony of many of its most thoughtful protagonists (Delandshere, 1994) will only cease when they settle for collections, and deal openly and ethically with the personal and social consequences of that choice.

Assessment of quality as moral action

Persig (1991) makes a strong link between morality and quality; in fact, to him they are synonymous terms.

He looks at the relationship between evolutionary structure and the metaphysics of quality, and shows that there is not just one moral system, there are many: In the metaphysics of quality there's the morality called the "laws of nature," by which inorganic patterns triumph over chaos; there is a morality called the "law of the jungle" where biology triumphs over the inorganic forces of starvation and death; there's a morality where social patterns triumph over biology, "the law"; and there is intellectual morality, which is still struggling in its attempts to control society.

Each of these sets of moral codes is no more related to the other than this dissertation is to the flip-flop circuitry which controls the computer on which it is typed. Let's consider this in relation to our bridge; its quality as a physical structure in the inorganic world was unrelated to its quality as part of the social life of people; just as that in turn was unrelated to its quality in that intellectual world that can conceptualise its probable long term effects on the environment, and hence on the lives of humans not yet living.

Further, there will often be conflicts between the static social morality that would hold the physical or biological or social structure stable, and the dynamic evolutionary morality that would move it onward:

> Intellect is going its own way, and in so doing is at war with society, seeking to subjugate society, to put society under lock and key. An evolutionary morality says it is moral for intellect to do so, but it contains a warning; just as a society that weakens its people's physical health endangers its own stability, so does an intellectual pattern that weakens and destroys the health of its social base also endanger its own stability(Persig ,1991, p168).

In a morality based on stasis there is no confusion; what destabilises the social system is immoral, is an act of inferior quality. Yet in a static-dynamic view of evolution this equation no longer holds. The central problem then becomes, in Persig's (1991) words:

> How do you tell the saviours from the degenerates? Particularly when they look alike, talk alike and break all the rules alike? Freedoms that save the saviours also save the degenerates and allow them to tear the whole society apart. But restrictions that stop the degenerates also stop the creative Dynamic forces of evolution (p228).

It would be easy to say that the actors themselves are aware of whether they are saviours or degenerates, but this is problematic. There may be cases of genuine manipulation, of intentional evil, but these are probably rare. Most choices are internally processed as the competition of two positives, not as the balance of good against evil. And even when the latter is the basis of the internal dialogue, the "evil" may often be a societally imposed value that from another frame of reference could be seen as positive.

In both cases, the actor must act on a sense of "rightness," of "necessity" that overrides choice. The actor, like the observer, simply cannot tell what the ultimate quality of the action will be, because the actor can never predict all the consequences of action. To claim that the ultimate test is whether the act is free of ego is to beg the question. Any act can be interpreted as ego-dominated, even acts of transcending the ego, which are designed to nourish the "super - ego."

Finally, we are left alone with our own sense of identity, our own sense of integrity. After all the agonising, all the reflection, we are finally left with a sense of the flow of life, with the flow of one particular life, of one particular relationship; with a sense of appropriateness that on the basis of static moralities is sometimes most inappropriate. And we do what we must do. This is the essence of evolutionary morality; it is the essence of what constitutes quality in the intellectual sphere; it is the essence of the meaning of quality in any assessment event in which a product or a person is the focussed element. It is a demonstration of what Churchman (1971) and Campbell (1956) call the heroic mood.

Quality products

Traditionally the problem of the relationship between quality and standard has been solved either by ignoring it, or by emersing it in semantic confusion: by fuzzing the boundaries, by assuming the two concepts are isometric, by ignoring the logical type error, by claiming that high standards are of course synonymous with high quality. And as it is self evident (within mythical discourse) that we can measure standards, it follows that we have measured quality.

What we have done is something much more damaging; by identifying standard with quality we have confined quality to the straight and narrow, and thus denied its very essence, which is to be found in its spontaneous deviation from the constraints of geometric efficiency. For the standard is a preconceived point (however practically unmeasurable) on a predetermined scale. It may indeed be used to describe a work of conforming excellence, but is quite incapable of recognising the nuances of diversity, the force of spirit that transforms articulate parrots into creative people. One of the characteristics of works of high quality resides in their difference, not of measure, but of style. Quality is perceived not in differences in kind, but its differences in difference; not in differences in length, but in variations of depth: in short, quality diverts us from the linear, takes us to a dimension orthogonal to the flat. "Quality, consciousness, and experience are separate words for what is one whole, as one lived-process" (Beittel, 1984, p110).

The essence of quality resides not so much in the aspects or characteristics with which we attempt to describe it, but rather with the relationships between those aspects, and the coherence of the whole gestalt that those relationships produce, and hence with the meanings that such coherence implicitly evokes. And as with all gestalts, it is recognised as such only within the milieu of its production, only against the culture that is its backdrop, only in terms of the event through which it emerges. As no two products in this material world can ever be completely identical, so must the quality that characterises them also differ. As that quality is multidimensional, and contains relational aspects, it is idiosyncratic to each product, as well as to the conditions of its production.

In general, discourse on quality is not amenable to that "better and worse," "more or less" description that is a prerequisite for any measure, and hence of any standard, or any categorisation. It is sometimes amenable to discourse, and to aesthetic response, and even to comparison in some of its aspects. And quality is amenable to change, both in its own meaning, and to the meaning it generates in relation to the product it relates to. Hence such discourse may indeed invite change in the product being discussed, and agreement be reached by some or all concerned (in that particular consensual event) that there has been a positive shift in quality.

Such discourse, such agreement or disagreement about quality, is itself a process of quality control, no less effective because it is collaborative, and no less effective because people disagree. As such it could provide another method of certification, as indeed it more or less does among the elite of any profession; a fact that for many would make a stronger case in this argument than any other. For example, the final educational judgment of this work is with two examiners, who may differ greatly in their opinions.

Standard products?

So what? If in measuring the standard we have denied what is essential in quality, does it matter? Lack of official recognition of originality, a little repression of creativity, is unfortunate but hardly crucial in the world order. Yet the other side of the coin may well be crucial in the order of the world. For what is involved here is not a single instance of non-recognition, but the very production over thousands of instances of the thinking person, of the learning person, of the person in work, of the person with authority; of, indeed, the moral, rational person.

For the standard is more than just one of many nudges and winks that lead the child to God. The standard, as applied continually through the strictures and structures of family and school and occupational work, at first externally and then through internal absorption and prescription, is the major mechanism, the quintessential carrot and stick, that moulds and shapes, that produces and creates that consciousness that defines the way each person sees the world, thinks about it, and acts within it. Not entirely, but largely so. And the individual produced through the notion of the standard, with its sharp cutting edge of adequacy, is a much more conforming, accepting, black and white, uni-dimensional person, and hence one far more socially controllable, than is one produced though the more spontaneous, multi-dimensional and unpredictable notion of quality.

Maybe we don't need to de-school. Maybe all we need to do is to acknowledge the arbitrariness and error that permeates standards and their measurement, extol the virtues of immeasurable quality, step lightly and quickly aside, watch the catagorisation structure crumble, have faith in chaos theory to articulate another structure, and hopefully nudge it in the direction of greater rationality and equity, truth and compassion. But that's another story.

Summary

The notion of the standard intervenes in the discourse about quality, and severely distorts it. The standard is a member of the class quality, is separated from it because of properties of measurement accuracy it is purported to have, yet is still confused with it. When the standard is seen, realistically, as unable to perform its function, we must return to quality as the notion with sufficient mythical, ideological, and intellectual status to replace it. This would predispose us to a rather different political structure, and to the recognition of a world in which simplistic notions of linear competition and dichotomous categorisations are replaced by more complex, ecological, and collaborative axioms.

Part 4: Error analysed

Chapter 13: Four faces of error
Chapter 14: What do tests measure?
Chapter 15: The psychometric fudge
Chapter 16: Validity and reliability

Chapter 13: The Four Faces of Error

Synopsis

The meaning of error in each frame of reference for interpreting assessments is now considered: In the Judges frame the phrase "error in the Judge's frame" is recognised as an oxymoron; in the General frame error is conventionally defined in statistical terms that ignore or underestimate some of the considerations, and the unattainable true score is seen to be a theoretical construct that need not relate to any external reality; errors are hidden in the Specific frame, and some of the Pretenders to this frame, namely mastery tests, criterion referenced tests, and competency standards, are briefly examined; finally in this chapter the meaning of error in the Responsive frame is considered. As this frame involves human interaction and discourse, error is what disrupts or disturbs movement towards clarification of meaning.

Assessment discourse is necessarily confused and confusing when the frame of reference within which the discourse is occurring is not specified, or when it involves definitions and methods where the actual frame being used is misrepresented.

The meaning of error in different frames

As soon as assessment data are committed to paper, their material permanency is dramatically increased. Likewise, the span of their associations is spread and emphasised. No longer just a description of a particular performance, the assessment becomes interpreted as a measure of knowledge and ability, an indicator of achievement on a course of study, and a predictor of future success or failure. Participation in an event has been transformed into an attribute of a person.

To estimate error is to imply what is without error; and what is without error is determined by what we define as true, by the assumptions of the frame of reference that forms our epistemological base.

There are four, at least, frames of reference for assessment. Four different sets of assumptions about the nature of the exercise. So within each of these frames the meaning of error, as defined by the assumptions of that frame, is different. Just as the meaning of error within each frame will be different again if judged by the assumptions of another frame. It is these differences that will be examined in this Chapter.

Error and the Judge

The Judge assumes omnipotence and infallibility within limits. The limits are defined by the particular performances with which the Judge is presented. These are the facts of the case. The task of the Judge is simple. He examines the performance of the accused, in whatever form it may be presented, he relates this performance to the standard, and then describes it accordingly.

He does this without error.

So problems that relate to error such as labelling, construction, stability, generality, prediction, categorisation, values and distortion of learning are, to the Judge, irrelevancies. For Judges are practical people, concerned with the realities, with what is, rather than what might be. And for them reality is the answers written on paper, is the art poster presented, is the motor repaired; in short, is the performance or artefact with which they are presented.

Questions of ability and stability, of looking to the past or to the future, are both irrelevant and unsettling. Irrelevant because they are outside the limits of their scrutiny. Unsettling because they trigger notions of a subject.

What sort of jargon is that?

Is what?

Trigger notions of a subject, for God's sake!

You find that a bit obscure?

I find that absolutely obscure.

I was alluding to the difference between subject and object.

I'm none the wiser.

An examination paper is an object. A grade is an object. A standard is an object. The Judge relates these objects. And he claims to do it quite objectively. A computer, programmed correctly, would also do it objectively. Objectively in this context means that the process is purely rational, untainted by emotion or expectation of any kind. The Judge is firmly positivist in his stance; he rationally assesses what is out there in the real world to be described.

Seems eminently reasonable.

Indeed, if somewhat inhuman. An observer in another frame of reference might see the Judge as myopic and deluded. He might see the Judge immersed in a totally subjective world triggered by the statements, now confined to paper of the person being assessed. Further, he might see the comparison with the "standard" as an intuitive rather that rational process, affected by images, emotions and expectations stimulated by script, time and style of the answers as much as by content.

> *That also seems eminently reasonable.*
>
> *Regardless, it is necessary for the Judge to deny such subjectivity in order to maintain the role of impartial expert, of perfectly calibrated measuring instrument. The Judge considers his work as objective, and so is unsettled by the notion of the subject, the four dimensional person who is assessing, and the four dimensional person who is being assessed*

Most teachers marking tests and assessing student work, and most public examiners, work within this frame. So most educational assessment is, by definition, error free.

Sometimes it is necessary, because of numbers of students, to have more than one Judge. There may be a number of Lesser Judges and a Chief Judge. In such situations it is accepted that ratings from lesser judges could contain some error, of the order of one or two marks in a hundred. To minimise this possibility, sample answers for questions might be prepared, with detailed marking shedules.

Sometimes a further check is made of papers just one or two marks below the cut-off points for failure. The Chief Judge will examine these to ensure that there has been no error, thereby restoring the myth of infallibility.

Reducing error in the Judge's frame of reference is not a problem. There is no error, except in the special cases of Lesser Judges and crucial decisions. In that case the error is the difference between the original assessment, and that of the Chief Judge.

Note that the Judge is infallible regardless of the form in which he presents the assessment. He may compare with the standard in any way he thinks desirable. The Judge is perfect in his rank orders, scores, grades, or other normative classifications. He is equally impeccable should he present his assessment in any other form, such as verbal description, moral tirade, or hologrammed logo.

The important point to understand is that the Judge is part of a social and political structure in which the inviolability and accuracy of the Judge's decisions are crucial elements. To suggest that the Judge may be in error threatens the stability of that structure and its accompanying mythology, so it is an act both treasonable and blasphemous: treasonable because it undermines the structure of society; blasphemous because it denigrates one of its icons.

In the hundreds of letters I have read in newspapers complaining about examinations, I have never seen one that suggested that the Judge, because he is a normal person, may make whopping big errors! So to the general public the Judge is not a normal person, and makes no errors.

Error and the General

Most of the book space and discourse time about this frame has been appropriated by those associated, corporately or academically, with the test construction industry; by those who produce and sell achievement and ability tests of many and varied kinds. Or by those who play in a scholarly way with mathematical models that might be used by those who construct such tests. (Nairn 1970). I shall deal with this world specifically in Chapter 15, the psychometric fudge.

Within this frame as constructed by psychometricians the error is the difference between the true score and the estimated score

However, the logic of the frame does not require such elegant and complex mathematical manipulation. The mathematical models have, overall, been counterproductive. Their theoretical elegance has hidden their inapplicability to most practical learning and teaching situations; the mystification of their statistical constructs has hidden from teachers, students and public alike the enormous extent of rank order inaccuracies and grade confusion, and the arbitrary nature of all cutoffs and standards.

One further point needs to be emphasised here. The General frame contains no notion of Standard. It is about creating stable rank orders of students. Anyone, anyone with sufficient authority that is, is at liberty to arbitrarily define a standard somewhere along that rank order. But a standard so defined is obviously a relative, not an absolute, division.

Error and the Specific

In this section we will look at error in the Specific frame in its purest form of specific behavioural objectives, as well as in its degraded states of mastery testing, criterion referenced testing, and competency standards.

In this frame there is only one correct description of performance, and that is the unambiguous learning outcome defined in advance. It is assumed that learning outcomes can be defined so clearly that there is no doubt about whether a student has, or has not, matched behaviour to objective. In such a situation there should be no problem with labelling error because there is no labelling. Each objective stands alone, pure and clear in its pristine self description; context, task and standard clearly enunciated. (Mager, 1962)

Construction errors are another matter. Whilst it is assumed in this frame that any outcome relevant to a particular course of study can be so specified, it is not claimed that all such relevant outcomes are in fact described. In some cases only those outcomes that all students are expected to attain are specified. Then we have a set of minimal learning outcomes. In asking "who makes this decision" we indicate a construction error. Why these particular objectives? And why these particular cut-offs for adequacy? It is apparent that behind the asserted certainty and objectivity of these objectives lies the usual minefield of idiosyncratic and arbitrary construction errors.

In other cases, a set of possible outcomes may be taken as indicators, and attainment of these is taken as evidence of achievement of related ones not directly assessed. And of course, no performance is ever a perfect indicator of a related performance, so hiding behind this wall of tightly specified objectives are all of the errors related to generality as well as to construction.

These construction errors, however, are all quite small compared to the massive one involved in the basic assumption of this frame: The assumption that any outcome pertaining to a course of study can be specified according to this frame; that all important outcomes can be specified in the form of a specific behavioural objective. In practice, it is just not so. This is what Messick (1989, p63) refers to as "construct underrepresentation".

This method of description is appropriate for situations where there are a finite number of tasks. Conceptually we are limited to tasks involving low level comprehension. As soon as we move into problem solving, analytic, application, or creative activities, there are an infinite number of possible task situations in which a student may be put in order to assess whether the student can demonstrate these more complex cognitive and practical operations. The tasks are limited only by the imagination of the test setters. And if we choose any one of these tasks, and describe them in such a way that they can be "taught" as a specific objective, then the task becomes one of low level comprehension. In other words, it must be a new task, a task previously unspecified, if these higher level performances are to be indicated (Bloom, 1956).

A student may attempt the task on a number of occasions if necessary, so usually irregularities in the performance of a particular student are not considered significant; unless, of course, a requirement of regularity over time is built into the objective. So errors in the temporal dimension are not applicable - unless, of course, we wish to infer that because a student has done the task, the student not only can do the task now, but on all occasions in the future. Such inferences are often made, of course. And they are utterly indefensible.

Prediction errors for an individual objective are enormous. But then, a specific objective does not claim that it would alone, or even in conjunction with other objectives, predict anything. On the other hand, as soon as it starts to describe itself with other adjectives, such as minimum, or essential, then it does open the way to predictive estimates of error.

Error and the Responsive

In the Responsive frame for any student there are many descriptions that are accurate and adequate to a particular purpose. Adequacy means that the description conveys sufficient information to carry the intent of the assessor and/or assessed into effect.

In this frame there is no competitive element, nor are the outcomes predefined in detail. Rather the assessor responds to the situation in terms of a particular purpose, which might be to describe how the student could improve the performance the next time (descriptive assessment). Or a responsive assessment might lead to a student's involvement in planning and assessing a course about maintaining a tractor (work required assessment). Or a responsive assessment might involve sharing a personal non-judgmental response to the student's work (detailed audience response).

While sometimes the criteria used for a responsive assessment might be preconceived, this is often not the case. The criteria emerge out of the totality of the situation, and so depend on the assessor's sensitivity, empathy and sense of quality. In addition, notions of adequacy are in general accepted for the subjective entities they indeed are, so become notions for considered opinion and discussion, rather than pretending to be absolute, accurately measurable qualities.

Responsive feedback then is part of a communication process which involves observation or other sensory input, interpretation, and response. It may in addition involve ongoing dialogue. Inaccuracy, in the sense of misinterpretations or misunderstandings may occur at any of these stages, as may obfuscations, denials, irrelevances, or contradictions. Empirically, this reduces to differences in interpretations, and there is no necessity in most cases to assume that there is some "true" interpretation or description. The aim is not to accept or reject the other's meaning, but to understand it.

In this frame, the person being assessed is also a potential observer and assessor, so self assessment can be an important part of the process. The communication process tends to be self-correcting, as the parties to the interaction both are concerned to clarify and understand what is being communicated. Accuracy then is concerned with the clarification of meaning, and error is reduced through openness of the communication channels.

Adequacy can only be determined by consequences. That is, to the extent to which effect conforms to intent. Again, error is reduced in as much as the assessed can feed back to the assessor the effect of the assessment, so that modification either of the description or the purpose can occur if necessary. This assumes that the assessed is aware of the purpose of the assessor's comments, and has reflected on their effects. So the continuity of open communication is as necessary as its initiation.

Keeping all communication channels open is of course more easily said than done, particularly in the social milieu that pervades most teaching-learning situations. For optimum reduction of error in this frame, both teacher and student would need to value openness over protection, autonomy over control, uniqueness over standardisation, complexity over simplicity, and tentativeness over certainty. In addition, each would need to be conscious of the potentially debilitating effects on open communication of the hierarchical structure in which their relationship is probably embedded.

More importantly, each would be wise to be aware of the potentially destabilising effects of their open communication on that structure, and of the social risk involved in such radical activity.

Summary

As the meaning of error changes with assessment mode, so do the methods designed to reduce such error. From a perspective of oversight of the whole assessment field, this is itself yet another source of confusion and invalidity, particularly as it is rare for any practical assessment event to remain consistently within one frame of reference.

Chapter 14: What do tests measure?

Preview

In this chapter I discuss in more detail the question of what it is that a test measures. In what sense can it be said to measure knowledge or ability? To what extent does it perform a ritual task and measure nothing? Or is it the wrong question? Should we rather ask, what do tests produce?

Tests and scales

A measure, or scale, assumes of course that equal intervals anywhere on the measure are in some sense of equal value. That the difference between sixty and seventy percent is in some way equivalent to the difference between twenty and thirty percent. So if a test is a measure then it must be a measure of something, and we would expect equal differences to represent equal differences in that something.

We know that a ruler measures length and the unit is a metre. We know that a clock measures time and the unit is a second. We know that a balance measures mass and the unit is a kilogram. And relative humidity measures what fraction of the water vapour the air could contain at a given temperature that in fact it does contain. So this is a pure number. Nevertheless, it is a ratio of two quantities that do have units.

So what does a test measure? And what is the unit of measurement? Let's look at the unit issue first.

It is clear that there are no units. The measure is a pure number. Unlike relative humidity, however, it is not a ratio of two measures of absolute humidity which do include units. Again, this supports the idea that the numbers are not measures, but ordinal numbers - numbers that represent an ordering of some kind. Numbers that describe a position in a series. Numbers in this case that assert that some performances, or people, have more of "something" than do others.

At this point it is worth mentioning that the whole paraphernalia of normalising scores and otherwise fiddling with them has two purposes: One is to try to magick a linear scale out of an ordinal one by making various sorts of assumptions about the distribution of the "something" that is being "measured"; the second is to produce "measures" that are mathematically pliable, that are accessible to the manipulations and pleasures of mathematicians; that will, in short, turn a horse race into a profession (See chapter 11).

Cultural differences

Back to the problem of the "something" that is measured by the test. For the most part, Europeans and their colonial converts on the one hand, and the United States and their spheres of influence on the other, have different approaches.

To the Europeans it has never been a problem. Inured by tradition to a religious belief in the Judge, they have generally accepted the proposition that the test or examination measures whatever the Judge says it measures. The acceptance of this "fact" denies the existence of a problem. The Judge says that tests measure student achievement. Pressed further, he or she might say that student achievement is a measure of what has been learnt on the course of study being tested. The test is simply that part of the course where learning is demonstrated. And the Judge, who holds the mystical secret and truth of standards, is able to convert this demonstration into a mark which is the true measure of what is achieved.

As I wrote that last paragraph I was aware of how "right" it sounded. Like all religions, there is a plausibility in its logical circularity that is terribly enticing, a simplicity in it's self-evident truth that gives a deep sense of security. Articles of faith are characteristically immune to both the challenges of logic, and the intrusion of empirical data. To paraphrase Horkheimer and Adorno (1972), faith needs knowledge to sustain it, and thus pollutes knowledge in the act of attaining it (p20).

The Americans, whose religious tradition is democratic and competitive rather than monarchic, have little faith in particular Judges or, for that matter, Presidents. Which is not to say that they do not revere even more in compensatory manner the institutions of power in which these fallible humans are niched. Regardless, their tests must be free of the Judge's subjective idiosyncrasies, and pay due homage to the competitive individualism that is central to the American dream.

The problem of subjectivity was (mythologically) solved through the medium of the "objective" test:

> The major premise of the American system of social morality is that every individual should have an equal opportunity to compete for the prizes offered. . . that every contest be objectively judged, as impersonally as possible, with no favouritism, nepotism, or any other kind of ism. To make this objectivity evident, access to preferred categories should, wherever possible, be granted on the basis of scaled scores that a machine can handle. (Friedenberg,1969, p28).

This has the added advantage, of course, of being "economically efficient," another central tenet of the American dream.

So the "something" that the test measures is measured economically and objectively, but we are still left with the sticky problem of what this "something" is. For when the Judge goes away, this problem raises its (previously covert) head.

Over the years, American test gurus have come up with a plethora of things that they claim to be measuring; intelligence, specific ability, attainment, achievement, competence, factors of the mind, specific outcomes, curriculum objectives, minimal competencies, true scores, universe scores, latent traits. An interesting oscillation between physics and metaphysics, between outside behaviours and inside mind-potential, between performance and hypothetical mental structures. Be assured however that efficiency has been conserved. In many cases the same test item can be used to measure all of these "things." (Nairn 1980; Taylor, 1994; Sternberg, 1990)

The simplest conclusion is that multiple choice tests measure exactly what the people who construct them claim that they measure; the definition of the abstraction they claim to measure is simply the score on the test. Which puts the Americans in a similar position to the Europeans, with the substitution of test agencies for individual Judges, of an elitist junta for the monarchy.

One corollary of this conclusion is that the tests really do measure something but no one is sure what it is. In the light of all of the evidence this seems unlikely to me. Contradictions are predictable from the logical type confusions that are inherent in the whole test process.

A more plausible corollary is that the tests do not measure anything in particular, nor do they place people in any particular order of anything, except the order that participating in testing events of any sort tends to generate. But they do place them in an order, along a single line of "merit," and that is all they are required to do.

One more point is very significant. "Ability" or "achievement" tests like the Scholastic Aptitude Test do place groups of students (not to be confused with individual students) in an order very closely related to parental income and social class. In this sense they contribute significantly to the stability of an unequable social structure whilst at the same time producing an ideological smoke screen by asserting that they are ordering on the basis of individual ability. And the victim pays for the test. Fantastic! (Nairn,1980; Friedenberg, 1969, p29).

Social skills

In 1976 I was about to begin a five year research project looking at social development in school classrooms. At the time there was much educational discourse about teaching social skills, which many thought were in short supply in young people. "Improving social skills" was an objective in courses from grade one Mathematics to grade seven English to grade twelve Economics. As part of the preliminary work I visited schools in Australia, Canada and the United States, and talked to many teachers about the social development of their students.

These teachers were all interested in the social skills of their students. They taught young people from the age of five in infant schools to the age of seventy five in Ph.D programs. Yet in describing their students to me there was enormous similarity in their descriptions. It went something like this: "When they first come to me they are pretty bad. Inarticulate really. Stumble over words, tend to answer just yes or no. Can't put two coherent sentences together. Can't listen properly. Can't concentrate. Just don't seem to be able to relate to other people. Bad with their peers, and worse with me. Then as the year goes on and they get more practice in speaking up and their confidence grows they improve tremendously. By the end of the year I've generally been able to produce a class with quite mature social skills."

What particularly struck me about these conversations was that they appeared to be the same regardless of the age of the students. So how could the social skills of five year olds be the same as those of twenty five year olds?

Then I thought about my own experience over the previous two years as a "leader" of communication workshops; thirty teachers doing residential five day courses to increase their communication skills. Weren't they exactly the same? At the beginning of the week hesitant, not really listening to other people, insensitive to feelings. Then by the end of the week attentive and empathic, talking poetry rather than cliches.

Had we been asking the wrong question? Did this change have anything to do with learning new skills? Or had we, over the five days, changed the social environment so that it was now appropriate to engage in a different sort of dialogue? Had the group experiences produced enough trust and cohesiveness to allow for some flow in human relationships, to overcome the stultifying role restrictions and mistrust that characterise much of our normal discourse? Were these observed changes simply indications of emotional openness, with concomitant increase in divergent thinking and spontaneity?

The implications for our research were clear. The question we should address was not "How do we teach better social skills?," but rather "How do we develop the classroom group so that mature social relations and discourse are appropriate?"

How can a social skill belong to one person? At least two people are always involved, and what is appropriate interaction, whether verbal or non-verbal, must always be a function of the relationship between them, of the context of the communication. What appears to be a quality of the person, a skill, turns out to be a production of a particular environment, a particular aspect of a human interaction, a discourse appropriate to a social relation.

As with the quality of the bridge, so with the quality of social behaviour: Even if it can be labelled, the label can't be pinned on any particular object.

Knowledge

You rigged it.

What do you mean, I rigged it?

You wanted to prove your point about not pinning a label to a person. Then you chose social skills to talk about. And OK, you've got a case there. But what about intelligence? What about intellectual skills? What about cognitive achievement? What about mental ability? That's where the action is.

Certainly that's where the money is. Skills are what employers seem to want, and increasingly what education seems to be about. And as you suggest, cognitive skills, facts and knowledge and understanding, are at the high status end of the skills spectrum. But why are they so different to social skills?

Because they surely do belong to a single person. You don't have knowledge in relation to someone. Analytic ability is not a relationship with another person. Reasoning skills are surely inside the person and not in some mystical relationship that characterises an event.

So let's look them in turn in more detail. Let's take knowledge first. If it's knowledge we're talking about, then it's got to be knowledge about something. So choose something.

Computers.

So how would you know that you had some knowledge about computers?

I've used them at work for various things; cataloguing, letters, drafting. So I know what programs to use for particular purposes, and I know how to use them.

In other words, you would reflect on particular interactions that you have had with computers, and on the results and feelings associated with those interactions?

I suppose so.

And you would interpret that recall of those experiences as knowledge?

Well, if I hadn't had the knowledge I couldn't have done the work.

But you just told me that you only knew that you had the knowledge because you had done the work.

Yeah, well that's now. But what about the first time?

What about the first time?

The first time I must have had the knowledge first or I couldn't have used the computer properly.

Tell me about the first time. Did you use the computer properly?

Well, you know. I had to mess around and experiment a bit before I got it right.

So the first time you had some knowledge, but not enough to do it properly?

Yeah.

And how did you know that you had enough knowledge to even make a beginning?

Well, that needed a bit of confidence, and a bit of taking a risk.

So it required a certain emotional state as well as a little preliminary knowledge?

Yeah, that's right.

And how did you know, or suspect, that you had that preliminary knowledge?

Well, I'd done some other work with computers. And of course there was the instruction manual.

In other words, you recalled other experiences with other computers. And you followed the instructions in the manual.

So is the knowledge in the instructions?

The instructions are meaningless without an event involving an interpreter and a computer.

Ok. If I had to follow the instructions then I didn't have enough knowledge. Reading the instructions became part of the event and enabled me to proceed. Now they are part of my experience that I can recall for future events.

So knowledge, once again, becomes, or at least involves, the process of recalling prior interactions.

So you reckon my "knowledge" of computers consists of reflections about real past events, or following instructions to produce an event which I can recall, in which I interact with a particular computer in particular ways. Knowledge appears in this case to be the construction, or the reconstruction, of an interactional event, a relational experience. Knowledge also implies that the emotional tone of that event is positive.

Exactly. Knowledge isn't something that you have. It's something that you do. It's something that is reconstructed in the present from memory traces of things that you've done before. You can carry out those reconstructions visually or in language in your own head, or in action with whatever objects are involved. And so knowledge of a particular field is continually created and recreated in the processes of selecting and applying memories of experience in that field.

Stories

L*et's make a slight diversion here to consider how this process of learning occurs developmentally in young children.*

All children, roughly between two and a half and four years of age, start to comprehend and make up narratives about their own lives.

> Also, adults of all cultures express their history, beliefs, values, and practices in the form of stories as psychological narratives. These stories are among a culture's most potent forms of self-expression and among its most effective forces for perpetuating itself (Stern, 1991, p133).

By creating a story, we create a reality. And we have as many realities as we create separate stories about ourselves in the world. It is in the creations of such stories that we define ourselves to ourselves. Out of our past we select and choose the experiences, with appropriate perceptions, that sketch the outline, and then fill the substance, of our stories. The firmer the story line becomes, the more selective our experience, and the more distorted our interpretations are likely to be, to maintain the story line. All this is fine, so long as we keep reminding ourselves that we are much more than our stories, that our experience is much richer than our perception and interpretation of it, and that the world is much more than our experience of it.

Yet there is another trap more subtle still. For not only do we get caught up in our own stories, we also get caught up in the stories of other people, particularly those we admire, or love, or are controlled by. For we do not live alone. We are social animals, and our life stories require other people to bring them into being.

Thus our stories about ourselves in the world are constructed out of our experience in the world. And this experience may come to us by direct involvement in the world, or involvement through the incorporated stories told us by others. And once these stories become accepted by us, they become part of our reality, part of our way of living in the world. Then we tend to construct our experience out of our stories. This is not a cause-effect relation, but an ecology of effects; our consciousness of the world, our way of being, involves an intimate interconnection of our experience, and the stories we use to make sense of that experience.

Our knowledge of ourself is just that interconnection.

Knowledge of a field

In just the same way do we construct knowledge in a particular field of study. We create events around the object of study, observe what happens, and then make up a story about what is happening. Or more likely accept someone else's story about what is happening. For any field of study is just such a consensus story, comprising what Foucault calls a "regime of truth." Then we use the story to help us make sense of other events involving the object, or other objects in that field.

This is equally true whether the field of attention is immense, as in mysticism or physics or history or engineering, or is small, as in building a table or washing dishes or driving a car.

So our knowledge of the field consists of descriptions of events involving a selected set of data constructed out of the relation between story and experience, between hypothesis and interpretation (possibly involving measurement), between conception and perception. As Wolf (1991) expresses it, "sophisticated thought follows a 'zig-zag' course between craft and vision"(p41).

But again, let us be clear on this fundamental point. The data, the knowledge, does not belong to the object of study. It is not a property of the object. Nor is it the name or a measure of a property of the object. It is rather information about the relationship of the object to its environment during a particular event, a particular interaction, suggested by the story in which it has a part to play.

Messick (1989a) comes close to this but does not follow it up. In claiming that tests "do not have reliabilities and validities, only test responses do," he goes on to say "that test responses are a function not only of items, tasks, or stimulus conditions but of the persons responding and the context of measurement" (p14). In my terminology, they are functions of events.

We could generalise. All knowledge is knowledge of the relations that identify events. And as we are observers at some point in the interaction, either at the level of direct observation, or at the level of constructing and interpreting the story that is the basis for the data collection, then we ourselves are involved in the interaction, and are thus part of the knowledge. And for the very reason that we are part of the knowledge, we are not that knowledge, and the knowledge is not part of us.

Human ability

In the light of the above, how are we to make sense of the notion of human ability, of capacity, of intelligence, of cognitive achievement, of some factor of the mind, of a latent trait?

These are normally considered properties of the person, attributes of an isolated mind, functions of an individual human consciousness. Yet our analysis of how we collect information about the other, or even how we obtain knowledge about our self, denies the possibility of such separation, and acknowledges the possibility only of information about relations.

I described knowledge of the field as a selected set of data constructed out of the relation between story and experience. Such selection is always in a context of some action, even if the most recent action is talking to oneself. Ability is a redundancy concept that acknowledges the action and then claims responsibility for it. It is an example of the common epistemological error of attributing a cause to the relational balance of an ecological system.

Semantically, this is achieved through the simple trick of nominalisation; of changing a verb into a noun, and thus of converting a process into an object. It is very simple: I do something, I am part of an event. Therefore, the causal logic goes, I am able to do the things I do (before I do them), otherwise I wouldn't have been able to do them. Therefore I must have (here comes the nominalisation) an ability, some property located somewhere within me, that allows me to do this thing that I do.

This is an example of the dormative principle. Keeney (1983), explains how it works:

> To invent a dormative principle, begin with simple descriptions of

> the phenomena to be explained. For example, a person may be described as unhappy and unwilling to work or eat. These descriptions can be classified as a category of symptomatic action such as 'depression'. The claim to then 'explain' these particular descriptions as the result of 'depression' is to invoke the dormative principle. What one does, in that case, is to say that an item of simple action is caused be a class of action. This recycling of a term does not constitute formal explanation.(p33)

The fact is the action: I run, or I try to run and cannot. What happens when the "ability" construct is introduced into the story? Now the reason I can run is that I have the ability to run. My running has a cause. I have some permanent property, some palpable attribute, that accounts for my running. My running is no longer a dynamic process of relationships between muscular and visual coordination, of memory and environmental feedback. My running is no longer a variable dynamic. It can be described as a causal relation independent of time.

My running is now explained by a little permanent stable packaged bundle of something inside me called "ability to run." It is a fixed static. As such it is a glue that helps fix me in time and space. It enables me to be compared, labelled and classified in terms of this property. It becomes part of my individuality.

What difference does it make? It makes world of difference, and a difference in the world. If the limits to my occupational choice and political power are largely determined by my cultural experience, by my practise in the field in which my interest lies, then most people might legitimately claim the right to such experience.

On the other hand, if my ability severely limits my possibilities in that field, then I have no legitimate claim to further practise. My exclusion is legitimised. I cannot become a doctor or engineer or lawyer not because of lack of opportunity or experience, but because of lack of ability.

Foucault (1992), in two condensed epigrammatic passages, sums up the essence of this argument:

> The individual is no doubt the fictitious atom of an 'ideological' representation of society; but he is also a reality fabricated by this specific technology of power that I have called 'discipline'. . . . power produces; it produces reality; it produces domains of objects and rituals of truth. The individual and the knowledge that may be gained of him belongs to this production (p194).
>
> . . . the disciplines characterize, classify, specialize; they distribute along a scale, around a norm, hierarchize individuals in relation to one another and, if necessary, disqualify and invalidate (p223).

It is not by accident that whenever universal education claims to equalise opportunity to cultural immersion and hence occupational choice, at the same time examinations and psychological labelling provides upper limits previously applied through the mechanisms of class and caste. The basis of the highest morality of any society has always been the maintenance of stability.

Conclusion

So what does a test measure in our world? It measures what the person with the power to pay for the test says it measures. And the person who sets the test will name the test what the person who pays for the test wants the test to be named.

The person who does the test has already accepted the name of the test and the measure that the test makes by the very act of doing the test; when you enter the raffle you agree to abide by the conditions of the raffle.

So the mark becomes part of the story about yourself and with sufficient repetitions becomes true: true because those who know, those in authority, say it is true; true because the society in which you live legitimates this authority; true because your cultural habitus makes it difficult for you to perceive, conceive and integrate those aspects of your experience that contradict the story; true because in acting out your story, which now includes the mark and its meaning, the social truth that created it is confirmed; true because if your mark is high you are consistently rewarded, so that your voice becomes a voice of authority in the power-knowledge discourses that reproduce the structure that helped to produce you; true because if your mark is low your voice becomes muted and confirms your lower position in the social hierarchy; true finally because that success or failure confirms that mark that implicitly predicted the now self evident consequences. And so the circle is complete.

Chapter 15: The psychometric fudge

Synopsis

The first part of the chapter details some of the ways in which psychometricians fudge; by reducing criteria to those that can be tested; by prejudging validity by prior labelling; by appropriating definitions to statistical models; and by hiding error in individual marks and grades by displaced statistical data, and implying that estimates are true scores.

In the second part of the chapter a number of specific examples of fudging are detailed; in particular, the item response theory fudge, selection and prediction fudges and the great Queensland reliability fudge.

Constraining the definition

Reliability and validity are two concepts dear to the heart of test constructors and others involved in the field of psychological and educational measurement. I'll begin my analysis of the fudge that characterises the field by looking at reliability, or the lesser fudge.

Reliability in classical test theory is (indirectly) an estimate of the error you'd expect if the student did a hypothetical parallel test. And in generalizability theory it's an estimate of the difference between the "universe" score and the score on any particular test. In both cases it's about the reliability of the test, or more accurately of the test-testee interaction, and not of the assessment; of the extent to which two tests give the same score, not the extent to which this particular description of student performance, based on a test, confirms or contradicts other such descriptions, which may or may not include a test (Behar, 1983, p19).

Note the way the mathematical model simplifies and constrains the world. It would be easy to believe the reliability of the test was about the extent to which the test describes course outcomes or student performance or work successfully completed. It isn't. It confines itself to the closed world of the test. It's about its ability to reproduce itself.

Mathematical models and true scores

The concept of the true score or universe score is central to the derivation of the theory. That is, it is a theoretical assumption. That does not mean that it necessarily has any place in the interpretation of the theory, that it corresponds to some measurable property of real people. And even if it does, the theory indicates that we can never know the true or universe score, only an estimate of it. And that estimate is always associated with error.

So in practice, in the world out there, there is no true score that can be attached to a person or an event. There is no thin line beside which a number is placed. Even before the empirical evidence starts to come in, there is only a wide fuzzy band, and all we can say mathematically is that the true score is probably in there somewhere. And if it is only probably in there somewhere, then for all practical purposes, for an individual person it isn't in there at all. In practice there is no true score. There is no stable rank order. And if in practise there is no stable rank order, then there can be no stable practical standard.

The history of achievement testing represents an enormous confusion of theory with practice. A model is not true or false. It is useful in as much as its predictions accord with empirical data at some points. It is not necessary that the assumptions of the theory correspond to actual situations in the world in which its predictions are applied. The assumptions of quantum mechanics from which the theory derives cannot be validated empirically. That is why they are assumptions. The metaphor in which the assumptions may be enclosed is useful in as much as deductions from the theory are experimentally verifiable. But such assumptions are not considered "true." Nor are they considered as having some "real" existence out there in the "atom."

Psychometricians on the other hand assert that their assumptions about a true score or universe score imply that such a score refers to some attribute, some measurable property, of a person. The person can be then classified, because the number is a measure of something called achievement, or ability, or whatever. In Criterion-referenced tests it is achievement in a specified "domain" of knowledge, and is called a "trait."

Regardless, this achievement is assumed to be some psycho-cognitive state which can be accurately described by finding a corresponding point along a one dimensional scale.

Why are these very intelligent people wanting to insist that their theoretical assumptions are consistent with empirical reality, when theories in general require no such correspondence? And when the fundamental assumption, the primary axiom of this particular theory, is that such correspondence can never be achieved? Why this enormous urge to represent uni dimensionally a variety of human performances which are obviously multi dimensional? Why this obsession with numbers, this illusion of numerical accuracy, this delusion of descriptive adequacy?

At this time, let us merely note that all of these activities are related to a psychological ideological assumption about human ability, or skill, or achievement. Some particular quantifiable quality of people that belongs specifically to them, and is thus independent of gender, race and class; that is unsullied by environmental factors; that is a permanent fixture of the person independent of the conditions of its production. That is, indeed, the clinging legacy of the nineteenth century belief that "intelligence was a unitary and immutable trait. It had no kinds or varieties, only ranks."(Wolf, 1991, p36).

As well these assessment activities are related to an ideological social assumption that this quality may be quantified and be represented along a uni dimensional line of almost infinite length, along which each person may now be accurately placed and categorised, their place permanently fixed, and their relative position in the order of things firmly established. And this conception of "ranking, fixedness, and predictability provided the "scientific" basis for two enduring institutional responses to the diversity of styles, cultures and academic backgrounds of students: universal testing and the systems of tracking students." (Wolf, 1991, p38).

And, further to Chapter 4 , note that

> This portable cumulative record of individual worth and achievement is central to bureaucracy and psychology alike. . . the inscriptions in individuality . . . make the individual knowable, calculable, and administrable, to the extent that he or she may be differentiated from others and evaluated in relation to them. . . individuality has been made amenable to scientific judgement. . . With psychometrics the previously ungraspable domain of mental capacities has opened up for government. What can now be judged is not what one *does* but what one *is* (Rose, 1990, p140).

The General frame and the true score

The logic of the General frame does not require any notion of a true score. The true score is a statistical artefact, a mathematical artifice, devised to defend a quite fantastic and monstrous proposition about ordering and classifying with great accuracy large numbers of people. Here is that monstrous proposition spelt out in more detail.

The political proposition that is being rationalised, justified, mystified, constructed and implemented in the notion of a true score is this: that it is possible in any area of human achievement to produce an accurate order of merit of "ability" in that area, and to attach to each person a number, a score, that fixes them firmly in position within that hierarchical order.

What do we actually know empirically? That under certain conditions it is possible to increase the stability of the rank order of merit of people on "test" results, in "test" situations. And that the more we can eliminate personal idiosyncrancies of setters and markers by averaging, and the shorter the time span of repeating the testing, the more the rank order is generalizable to other setters and markers of similar tests constructed by similar people.

We do not know empirically whether there is an asymptotic limit to this stabilisation; theoretically, and practically, there is always an error of measurement. We do know that this fits empirical data quite closely in regard to sampling assessors for marking. That is, when students do very similar tasks and the idiosyncrasies of assessors are "averaged" out.

We do not know empirically whether a similar stabilisation occurs when results are averaged over different occasions. There is no a priori reason to believe that they should be, especially for achievement tests with a high memory component. Indeed, there is every reason to believe that the actual performances of particular students would vary considerably, and differentially, when assessed over time, given that their forgetting curves are non linear and of different shapes. Thus sampling across these dimensions could produce an increase in error in the General frame, not a decrease. It would be very dangerous to collect such information, however, for it would contradict the assumption of stability that the notion of skill or ability implies.

Empirically the true score is not known, and can never be known. Empirically estimates of the true score can be obtained, and these are always different, because all of the measurements we make contain an error. In practice then, error is indicated by the difference between estimates, not between estimates and some hypothetical "true score." That is why the notion of true score is not necessary for simple and specific and individualised estimates of error, though theoreticians and ideologues may well require the idea for their own particular purposes.

The notion of the true score, then, despite its enormous ideological importance, is practically unattainable, irrelevant, and misleading. It is a theoretical input to the mathematical theory of testing, not a practical output. The statement that there is a true score is a statement about a theoretical statistical assumption, not about an attainable empirical reality. Further, such assumptions of mathematical models need have no direct links to any properties or aspects or qualities of phenomena "out there" in the real world.

> Note that we do not define true score as the limit of some (operationally impossible) process. The true score is a mathematical abstraction. A statistician doing an analysis of variance does not try to define the model parameters as if they existed in the real world. A statistical model is chosen, expressed in mathematical terms undefined in the real world. The question of whether the real world corresponds to the model is a separate question to be answered as best we can (Lord,1980, p6).

Lord then agrees with me, at least on page 6. More of Lord later. For now, having seen how the fudge about the true score works, we'll examine some of the others. One really big one relates to test items.

Models and items

There is no doubt that one way to get information about achievement (what a person has done), or skill (what a person can do), or ability (what a person could do given the opportunity), is to get them to answer some questions about what it is they are supposed to have achieved or have the ability in. And one rather contrived way of doing this is to use pencil and paper tests. Further, a particular method of this technique is to use test items of a multiple choice or short answer form.

It requires an enormous suspension of rational thinking to believe that the best way to describe the complexity of any human achievement, any person's skill in a complex field of human endeavour, is with a number that is determined by the number of test items they got correct. Yet so conditioned are we that it takes a few moments of strict logical reflection to appreciate the absurdity of this.

Test items not only determine the form and media of testing as paper and pencil tests, but also specify the type of question as short answer or multiple choice. In other words, talk of test items tends to narrow dramatically the sort of performance situation in which the person being assessed is to be put, and also severely limits the sort of description that might be given.

Why is this important? Because psychometricians have defined reliability and generalizability in terms of test variance, which is in turn determined by the characteristics of test items. Likewise, estimates of construct validity, on the rare occasions they are estimated empirically, are determined by statistical manipulations of item characteristics.

By appropriating terms like reliability and generalizability and validity, and defining them in terms of the mathematical properties of particular tests, professional test agencies and examining institutions perpetrate another grand fudge. These concepts become narrowly construed as properties of tests, or relations between numbers, rather than as useful criteria on the basis of which concerned people may judge the whole assessment exercise.

Item response theory and the absolute scale

Item response theory allows us to construct a scale in the same way that classical test theory and generalizability theory enables us to construct a true or universe score.

The magic is in the word "construct." It is theoretically constructible, not empirically constructible. In fact, the theory determines that the scale is absolute but improbable; the actual scale produced measures the probability (or if you prefer, the improbability) that any person to whom the scale is applied actually has that reading on the (theoretically) invariant scale that the theory constructs.

Just as objective tests are highly subjective instruments in which the marking can be done objectively, but it is implied that the assessment is objective; and just as the true score can never be measured but it is implied that the estimated score is that score; so the invariant scale of the criterion referenced test can never be physically produced, but it is implied that the test produced contains that scale, rather than its very error-prone physical manifestation.

Criterion referenced tests

Criterion referencing, as applied by professional test agencies, is not directly referring to course objectives or to student learning. Criterion referencing refers directly to test items. A criterion referenced test is one that is proscribed by tight delineations of the structure of particular tasks to be included in the test.

Advocates of criterion referenced tests often claim that the performance on such a test is judged in relation to an absolute rather than a relative standard. That is, that scores on criterion referenced tests are measures of achievement in a particular domain and do not depend on relative merit, but are informative in their own right.

This claim is another psychometric fudge. Criterion referenced scores are in no way absolute scores. They are norm-referenced. The norm-referencing is done prior to the test construction process at the item level, and not at the total test level during a specific application of the test. (Behar 1983, Glass 1978)

Criterion referenced tests contain all of the errors of Mastery tests plus one additional labelling error of great ideological significance. A sub-group of tests in this area, called sometimes Domain referenced tests, have developed a whole theory based on test item characteristics, which is very efficient. Efficient in the sense that students can be tested with less items than in the random sampling model for the same error (an error which, as usual, is never attached to individual scores). This is achieved by using known levels of difficulty of the items (based on random or other specified population estimates), in computing the student's score.

Nothing wrong with this of course. Except the labelling claim that these scores are absolute measures of a "latent trait." What is a latent trait? It is some "hidden characteristic" which some students have more of than others, and which is measured by the test. And those who have more of it are more likely to be able to answer correctly the more difficult items.

As all of the items in a Domain referenced test relate to some particular area of learning, such as reading comprehension, or computer skills, or simple calculus, or newspaper editing, or social skill, or whatever, then it doesn't really matter what "latent trait" means. The assertion that "it" can be measured absolutely is what constitutes its ideological power. Here is the ultimate rationalisation for intellectual and social stratification. Here is the number that describes each person's place on the continuum of ability or skill or whatever for any label that testing agencies wish to attach to the domain of items.

On the surface, of course, it is the specific label that assumes social importance. The claim being made, or at least strongly implied, is that such a test is an absolute measure of reading comprehension, or computer skill etc. But in focussing on the label, we are likely to miss the frightening significance and ideological sleight of hand that produced the "latent trait" as some substantive property or quality permanently attached to the person tested, somehow magically unrelated to the highly subjective, contrived, interrelational world where a student sits at a desk, reads some questions, and places ticks in computer marked boxes.

Such tests construct current fashionable truths. They are being presented as the latest panacea for testing human ability, or "skills" or "competencies" as they are now called; they are being presented as the theoretical support for an invasion of competency based assessments in all areas of human measurement (in schools, businesses, bureaucracies, or where-ever else hierarchies operate). So we should be clear about three things:

The first is that constructing a domain referenced test and naming it produces no evidence that the tests measures any sort of trait or ability that can be attached to an individual person (Lord, 1980).

The second is that they are not absolute, or error free measures; the scores are related to relative merit, and there is no "standard" performance or score that relates to any minimum or other grade of "competency" that can be theoretically attributed to any score (Glass, 1978).

Which takes us to the third point, which is a logical conclusion from the previous two. Domain referenced tests can make little contribution to a field of "competency" assessment which purports to describe (or more significantly measure) some "standards" of competency in various "skill" areas of human performance.

Limiting constructs, limiting error

Let's examine briefly how some of the more general criteria of assessment; labelling, construction, stability, generality, prediction, tend to be limited to what can be controlled by test makers.

Labelling is achieved by the simple act of giving a name to the true, or universe, or latent trait, score. Which means, in practice, to the estimated score. The errors implicit in the communication of what that label means, between those who define the course, those who teach it, those who produce the test, those who do it, and those who consume its product, are thus not considered. All of these people will give their various meanings to the label, and make their judgments accordingly. We may be certain that these meanings vary considerably. How much they vary will probably never be known, because it is not in the interests of any institution to uncover yet another source of error. Labelling errors are not currently considered in any estimate of test error. I believe they are immense.

If communication is its effect, then such confusions are, to the student, irrelevant. To the student the meaning of the label is the grade or the mark attached to it. Within the structure that contains the assessment system, the meaning of the label, as distinct from the meaning of the mark, amounts to little more than ideological gossip.

At least some students recognise the meaninglessness of the label. I remember vividly a television program which followed the fortunes of four students through the final months of their preparation for the University Selection Examination in New South Wales. One student in particular, a science student, a paragon, studied hard and reaped the ultimate reward. Straight A's.

Just after he received his results he was interviewed for the last time. He was obviously pleased with his success.

> "I suppose," the interviewer said, "this will be very useful to you in the future."
>
> "The marks?"
>
> "The understanding. The knowledge."
>
> "Oh that. No, I don't expect that to be of any use to me at all. I'm going to be a lawyer."

Likewise, construction errors are not estimated; they do enter the theoretical psychometric definitions of validity, but are in practice neither measured nor estimated. The major task of matching objectives to assessment to performance is assumed entirely by the test maker, and most of the errors within this activity are also disregarded, as easily as the errors caused by differing forms of assessment, and use of media other than reading/writing, which don't fit the format of test items on paper, are disregarded. It is assumed that the test is indeed contracted, and the performance required by the student indeed matches, the objectives of the course, or the criterion definitions of the test. Sampling processes that are used, even in professional testing agencies, are at the best primitive, and at the worst nonexistent. This part of test construction is nicely described as an "art" rather than a science (Nairn ,1980).

One thing is certain though; no course has stated as its major, or even minor objective, the ability to answer a pencil and paper test in a given time under stress conditions. And why not? Surely this is the essential behavioural objective.

Stability becomes narrowed to test reliability, more accurately called internal consistency, an internal test measure that cannot take account of variation over time and place and assessors. Theoretically test-retest reliability is one form of reliability, but in practice such estimates are rarely obtained.

Generality becomes narrowly construed as related to the extent to which the test samples the universe of possible test items, or how well the item specifications cover the domain. Generality becomes a function of test items and is called generalizability. Generalizability ignores previous performance in different contexts, forms and media. It ignores all performance other than the purely cognitive response to simulated experience of a multiple choice or written form. It thus ignores all cooperative and all production modes of expression. It reduces human response to the act of recognising a "best" answer, to conforming adequately to some authority's view of importance, relevance and reality, or to answering someone else's question in a particular way.

And prediction becomes tied to numbers and test scores. In this psychometric world we are no longer concerned with the extent to which actual people are helped to function in differential social situations of great complexity. Prediction does not attempt to describe the relationship between a particular set of learning experiences for some person, and how helpful that is in some future situation for that person. Rather it ranks a group of people on their "success" in the "learning" situation, then ranks them again in some criterion situation. The correlations between the two rank orders represents the predictive value of the test. Not of the course, of the test. And not of its relevance to the quality of their performance, but to its correlation with some person's or group's ranking of their relative performance. And note that even if this correlation is high, which is unusual unless a similar test has been used to measure the criterion, this tells us nothing about whether the relation is in any way causal.

How the fudge works

The psychometric fudge occurs through the following processes:

Firstly, the criteria by which assessment is determined are chosen so that they are easily adaptable to the construction of tests and to the statistical manipulation of test data. Criterion-referenced tests are just that: Only those criteria that are appropriate for referencing test items are chosen.

Secondly, the validity of the test is prejudged by labelling it to describe what it is supposed to measure. Such is the power of labelling that this exercise in wishful thinking, this untenable assertion, is interpreted by most people, including the test constructors who become entranced with their own propaganda, as being an accurate description. At a deeper level still the mathematical theory itself contains such terms as true score, ability, and trait before any empirical information at all is available; that is, before any connection (let alone correspondence) with the world outside mathematics is established.

Thirdly, definitions are appropriated and defined to fit specific statistical models; in particular, by narrowing the universe of possible test situations to a universe of possible test items (random sampling model), or by narrowing the universe of possible test items further to the universe of suitable test items (domain referenced testing). In both cases the performance of students outside of such test situations is disregarded, or downgraded, and the right to appropriate the personalising labels (ability, trait, true score) is assumed.

Fourthly, the data is presented in a way that is misleading at best and deceitful at worst, by hiding error of individual marks and grades with obscure and displaced statistical data, thus implying, to all but the statistically sophisticated, that estimates are "true" scores. Further, the implication is made that such tests are accurate as predictors, claims that in most cases cannot be substantiated (Reilly, 1982). Finally, estimates of confusions and errors related to construct validity are ignored, usually theoretically, and almost always practically.

We could look at these fudges as things done by individuals, and thus attributable specifically to them. From this psychological frame how could we make sense of this fudging behaviour? At best the fudges can be interpreted as logical or psychological slips propped up by delusions of grandeur. At worst they represent academic chicanery and political manipulation in high degree (Nairn, 1981, p58).

If we regard this in a sociological context, however, a different picture emerges; psychometricians may well be regarded as the moral guardians of the age of competency, the high priests who hold society stable by propagating, preaching, and propping up the gospel of the Standard, and the cult of the linearly determined individual that it constructs and supports.

In the beginning

"What's in a name?" Bill Shakespeare said, "that which we call rose by any other name would smell as sweet." Maybe so, yet that which we call a trait when it is just a mathematical function takes on a different odour indeed. Names have a magic of their own, and the stickiness of the name is very dependent on the power of the namer.

Lord (1980) produced the seminal work on item response theory, in his book Applications of item response theory to practical testing problems. It is possible here to trace in detail the birth of a fudge.

Early on there are some laudably honest statements:

> True score theory shows that a person may receive a very low test score either because his true score is low or because his error score is low (he was unlucky) or both (p5).
>
> The true score is a mathematical abstraction. A statistician . . . does not try to define the model parameters as if they actually existed in the real world. A statistical model is chosen, expressed in mathematical terms undefined in the real world. The question of whether the real world corresponds to the model is a separate question to be answered as best we can. It is neither necessary or appropriate to define a person's true score or other statistical parameter by real world operational procedures (p6).
>
> In item response theory . . . the expected value of the observed score is still

> called the true score (p7).

Admittedly, our laudability quotient diminishes as we reflect on the use of the word "true." In what sense can it be true if it doesn't exist in the real world? Why call it true if it can't be measured. But perhaps it is true in a mathematical sense because it is a necessary conclusion for the premises of the theory? Not so, it is merely the name of a variable assumed in the theory.

Undeterred we press onwards. Five pages later Lord commences the serious work in developing the theory:

> Let us denote by ø the trait (ability, skill, etc) to be measured. For a dichotomous item, the item response function is simply the probability Pø of a correct response to the item. . . it is very reasonably assumed that P increases as ø increases (p12).

Now this is truly remarkable paragraph. The word "trait" has not appeared before. Where did this "trait", this "ability", this "skill" come from that is being measured? What does it mean? Lord "very reasonably assumes" that as this thing increases, the probability of answering a particular test item increases. But why do we need this thing at all? And why is it named a trait or a skill or an ability, which are hardly "mathematical parameters"?

We wait expectantly till page 45 to find out what ø means mathematically. "A person's number right score . . on a test is defined . . . as the expectation of his observed score x. It follows immediately . . that every person at ability level ø has the same number right true score." Then on page 46 the crucial point finally emerges "true score . . . and ability . . . are the same thing expressed on different scales of measurement. " And just in case you missed it, the best estimate of this true score, this ability, is the number of items answered correctly on the test.

Thus on his own admission Lord has done exactly what he claims statisticians do not do. He defines the parameter as having "real world" status when he calls it ability. (Just as he infers it has some objective or propositional reality when he calls it true). Its mathematical status is simply the number of items answered correctly under the idealised conditions specified in the theory. It's empirical status is the actual number of items answered correctly, or some statistical manipulation of that number.

There is one more aspect of this fudge that we need to look into. It is the fascinating use of the adjective "latent" in front of trait. Hambleton & Swaminathan (1982) elucidate:

> Any theory of item responses supposes that, in testing situations, examinee performance on a test can be predicted (or explained) be defining examinee characteristics, referred to as traits, or abilities: estimating scores for examinees on these traits (called 'ability scores'); and using the scores to predict or explain item and test performance. . . Since traits are not directly measurable, they are referred to as latent traits or abilities. Any item response theory specifies a relation between the observable test performance and the unobservable traits or abilities assumed to underlie performance on the test (p9).

Of course, this is not quite true. Item response theory does nothing of the kind. It assumes certain characteristics of test items, and then generates a total score which is an estimate of the true score. Under certain conditions, "we can think of ø as the common factor of the items" (Lord, 1980, p19). The true score can only be guessed. The mathematical theory tells us the probability that it lies somewhere within a certain range of scores. Latent means hidden or concealed or potential. What is hidden, what is latent, is not any characteristic of the person, but a characteristic of the measurement itself. The examinee has performed, has participated in the event of answering test items. Nothing hidden or latent about that. So why the displacement? How did a latent measure become a latent trait?

Item response theory doesn't need any assumption about traits at all. The talk of traits and abilities is redundant and gratuitous. After all the terribly refined and elegant statistical manipulations, Item response theory simply produces a total score which (given knowledge of the structural characteristics of individual items) allows a prediction of the probability with which any particular item will be answered correctly by a person with that total score. It does require a certain consistency of correct (or incorrect) response for specific items on the part of the examinee. All else, as far as item response theory is concerned, is fantasy.

Incidentally, such prediction is in no way an explanation; to assume that is to evoke the dormative principle; the total score is just a summary of information about a particular person answering the individual items. Such a score cannot now be used to explain why the items were answered correctly.

On page 55 Hambleton and Swaminathan (1982) come clean; rather by accident that design, I fear. "Ability", we read, "is the label that is used to describe what it is that the set of test questions measures." Precisely. And what it measures is an estimate of probabilities of answering certain test items correctly. To what extent that measure relates to any "characteristic" or "trait" or "ability" of the examinee may only be known after "construct validation studies . . . (which) validate the desired interpretations of the ability scores" (p55). Shouldn't that read "validate or invalidate"?

Mistakes: probability, correctness, and checking

Item response theory cannot predict whether a particular person (whose true score we don't know but whose estimated score we do know), will get a particular item (whose characteristics we know), correct or incorrect. The theory will predict the probability of getting it correct. In practice it will either be correct or incorrect (probabilities are only 1 or 0).

So item response theory never even pretends to estimate what people know or can do. It only claims to estimate the probability that they can do certain things. Then the assumption (and that's exactly what it is) is made that this indicates an ability of the person in that area of cognition. It might mean something else. Or it might not.

When I worked as a test constructor I noticed one aspect of answering tests that was interesting. When groups of year 10 students did the 100 item tests most would finish in about ninety minutes. When groups of year 8 students did the tests most would finish in about 60 minutes. The year 10 students got slightly better results (about 0.3 S.D. better). Conventionally this would be interpreted as meaning that they had more ability, or simply more maturation. But given my perceptual data, perhaps it just means that they did more checking!

Psychometric selection myths and fudges

Hulin, Drasgow & Parsons (1982) complain that the controversy and rhetoric about standardised educational admission tests seem to have developed independently of the psychometric evidence about the usefulness of admission tests in reducing errors in prediction. They claim that Cleary, Humpreys, Kendrick, & Wesman (1975), Rubin (1980), Linn, Harnisch, & Dunbar (1981) among others, have produced summaries of large numbers of studies relating college and professional school admission test scores to performance in post secondary and postgraduate educational institutional institutions:

> The evidence is clear and consistent. Well-constructed tests of cognitive abilities are significantly and consistently related to performance in school. When appropriate corrections are made for restriction of range and other statistical artefacts, the validities of tests are appreciably large (p 281).

Claims such as this are very common. So on this occasion I thought I'd check out the references.

Cleary's (1975) data involved correlations between verbal and mathematical SAT scores on the one hand and High School grade averages and College grade averages on the other. The correlations ranged from 0.35 to 0.50. But the correlations between the High School and College grades were higher at 0.64. So two points about Cleary's study: firstly the correlations are at best only 25% better than pure chance. Is this "appreciably large"? Secondly, they were considerably lower than the correlations from grade averages, so why were they necessary at all?

Rubin's (1980) study involved the use of the Law School Admissions test to predict first year grades in 82 law schools. The correlations ranged from 0.03 to 0.5; after corrections for range (Linn, 1981), the correlations range from 0.2 to 0.7. In 14 of the schools they were below 0.35, which is 12% better than chance. Is this "appreciably large"?

When it is known that issues of construct validity introduce far more sources of error than are involved in simple predictive correlations of this sort, it is difficult to understand how this sort of justification, which is quite common in the literature, goes on for decades virtually unchallenged within the psychometric community; on the other hand, compared to the abysmally low correlations often obtained in such predictive correlational studies, perhaps they are appreciably large.

However, these studies raise another issue and another fudge; the correction (always upwards) of predictive correlations.

Fudging the predictive correlations

Correlations between a selection instrument and later performance are often corrected for range restrictions and for criterion unreliability. Range restriction is reasonable; generally some of the people tested were not selected, so had no opportunity to be in the final sample. It is considered appropriate by statisticians then to estimate what the correlation would have been had all of those selected actually been appointed. After the correction, of course, it is a correlation about something different; it becomes the estimated correlation between test performance and later performance of all those who sat for the test. Prior to the correction it was the correlation between test performance and later performance of all those who performed later. Different sample, different correlation. Which to use depends on what question you ask. Automatically raising the correlations is a fudge.

Correcting for criterion unreliability is a different matter. Most job tasks are multi-dimensional; that is, they involve many very lowly correlated tasks. And college grades are likewise composites based on lowly correlated components. If a single correlation is to be obtained a with multi-dimensional job performance the various ranks or gradings have to be collapsed into one single rank or grading; and that requires some arbitrary and explicit loading to be applied to each dimension (See Chapter 10 on Comparability).

Even when this is done (and it often isn't), there is still the assumption that there is indeed a meaningful rank order to be obtained. If most people in most jobs or in most courses do their work adequately (just as most people drive cars adequately), then we would expect correlations to be low, and ultimately, where training schemes are very adequate, to be zero. In such situations, the reliabilities would be low not because of rater inadequacy that can be corrected for, but because raters are attempting to separate performance when it cannot be separated, or/and are trying to pretend that a multi-dimensional performance is in fact uni-dimensional. In such cases it is obviously not appropriate to artificially inflate the correlations because of rater unreliability.

The changes are more than trivial. A study by Schmidt, Hunter & Pearlman (1981) involved 150 000 people, 2000 predictive correlations. Before correction the average correlations between eight aptitude tests and job performances in clerical job categories ranged between 0.15 and 0.25. After the statistical corrections, however, they magically rise to between 0.3 and 0.5. Still not good. In fact, still quite awful. But they certainly look better than before, and aptitude tests survive again to live another day.

The great Queensland reliability fudge

I was talking to the Principal of a secondary school in Queensland. Students in year 12 are assessed internally, with the help of some external monitoring. I suggested that there might be some problem with reliability. "It's 0.95," he replied with confidence. "Excellent," I responded with some scepticism. Then I decided to check the data.

The study is titled Random sampling of student folios: a pilot study (Travers, 1994). In this study

> . . . 1189 exit review folders of Year 12 student work were collected randomly from school subject groups across Queensland in December 1993

> and assigned to two hundred and forty review panellists in other districts. These exit review folders show the work of students who have received a result for that subject on their Senior Certificate. The role of the review panellists was to examine packages from schools containing ten folios, and for each folio decide a Level of Achievement and relative position in that achievement band (p 1).

The review panellists were given access to other marker's assessments and comments, as well as the school's assessment of the Level of Achievement. What they didn't have was information about the rung placements within each level of achievement (There are ten rung placements within each level of achievement) .

So this is not a blind reliability study:

> because it was not possible to reproduce all the conditions under which judgments about students were made by schools which supplied folios. In particular, panellists did not have the opportunity to observe student performance over an extended period of time as teachers do (Travers, 1994, p12).

The astute reader will already have noticed a contradiction here. The study was not constructed as a blind reliability study where no previous marks or grades were attached because they wouldn't have sufficient data to make valid judgments about levels of achievement. On the other hand they are being asked to make much finer discriminations regarding rung placements.

The astute reader will also doubtless have expected a very large halo effect, and would not be surprised if reliability coefficients, at least in relation to levels of achievement, were very high. As indeed they were. Eighty per cent of achievement levels remained unchanged, most of the aberrant cases being one level lower, indicating, no doubt, the "high standards" of the review panellists.

The overall correlation figure obtained for agreement between school exit and review level rung placements, on a fifty point scale, was 0.95. The authors were particularly pleased with the rung placement data:

> a rung difference of plus or minus one or two is not so much a significant difference as a demonstration of precision and accuracy . . . half the decisions about rung placement involved either assigning the same rung or one or two rungs lower. . . (this) suggests that not only do these panels read the folios very closely, but that they are able to arrive at decisions about standards that are both highly reliable and very precise (Travers, 1994, p17).

I did a little experiment. I listed fifty (hypothetical) folios in rank order of one to fifty, with ten papers at each level of achievement. Then, keeping them at the same level of achievement, randomly allocated new (reviewed) rung placements within each level. The rank order correlation was 0.95.

It follows that acceptance of given levels of achievement (halo effect), combined with random allocation of rung placements, is sufficient to account for the 0.95 correlation that was used to justify the whole procedure, not only of the pilot study, but indeed for the whole examination system, as evidenced by the Principal's comments.

Rather than evidence of precision in rung placements, which determine tertiary entrance scores, the data generates evidence of randomness, and another psychometric fudge is perpetrated by well meaning psychometricians on a gullible public.

The General frame and the true score

The General frame of reference as hijacked by psychometricians contains as an essential element of its assumptions the notion of a true score; a further element of those assumptions contains the notion that it is possible in some way or another to approach that true score; to get measures empirically closer to the true score by various procedures implied by the particular model. For example, in classical test theory by increasing the number of items on the test; in generalisability theory by sampling more tasks more randomly from a bigger collection of possibilities; in item response theory by having more items of appropriate characteristics which are uni-dimensional; in domain referenced tests by having the domain of items criterion referenced to a high degree.

Allied to this frame but not tied to it so tightly are the various notions of reliability and validity that have not been developed as part of the mathematical models mentioned in the previous paragraph, but have emerged from more general considerations of the notions of assessment, rather than of tests. In my terminology, these considerations have challenged the artificial constriction of the general frame by psychometricians, and have restored, through notions of construct validity and consequential validity, at least some of error components previously bypassed.

However, this has produced a contradiction with the notion of the true score that has not been made overt. For example, as described in Chapter 16, most achievement tests are not made more valid by increasing their reliability; on the contrary high reliability is seen to be, in most circumstances, an indicator of low validity. For most achievement areas involve a large number of disparate activities, and there is no a-priori, or even post empirical reason to believe that these activities are uni-dimensional, or otherwise closely inter-correlated.

I argue in Chapter 15 generalising the assessment events across contexts, or time, or media, or even value assumptions or frames of reference, does not (as does generalising across selection of test items or markers), reduce the standard error of the estimate; on the contrary, we have every reason to believe that it will increase such error, to a point where the whole notion of true score becomes unsustainable. After all it is not by chance that so much space is given in test manuals to ensuring the conditions under which the test is given are kept constant. Obviously this indicates the fragility of the test to contextual shifts. (On second thoughts, it could be as much a ritual designed to imply scientific accuracy, and sustain the notion of fairness). Regardless, it is clear that contextual shifts increase the error term, whilst contextual control artificially reduces it; artificially because no argument is ever given, nor could it be sustained, that this particular test context is superior to any other to the measurement of this "ability." So once again the price of higher reliability is lower validity.

Preview

We could go on dealing with the specifics, but it is time to present the greatest fudge of all. Validity. For as will become clear, the very definition of validity creates a discourse around it where every test may be assumed valid until proved otherwise, and as there are no specific descriptions as to how such a proof might be constructed, and no specific standards of acceptability to which such descriptions might be compared, all assessments may claim to be valid.

Chapter 16: Validity and Reliability

Preview

The professional theoretical face of assessment discourse asks the question, is the test reliable? More ethically orientated assessors ask the additional question, is the assessment valid?

The public wants to know, is it fair? And the more critical of them might add, are people being violated?

In this chapter some of the more recent work on validity is discussed, and its positioning as advocacy demonstrated.

Reliability is also discussed as a problematic, rather than as an obvious prerequisite to validity.

Validity

"Validity," states the first sentence of the APA Standards of educational and psychological testing (American Educational Research Association, 1985), "is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p9). It goes on immediately to explain that: "Test validation is the process of accumulating evidence to support such inferences."

Which all sounds very scientific and objective and devoid of bias. But is it so? Let me, from my own particular concern with the test taker, rewrite the first sentence to dovetail more accurately with my concerns.

"Invalidity," states the first sentence of the alternative tract, "is the most important consideration in test evaluation. The concept refers to the inappropriateness, meaninglessness, and uselessness of the specific inferences made from test scores. Invalidity or error estimation is the process of accumulating evidence to problematise and ultimately reject such inferences."

It should be clear even from this small rewrite that a text that began with the second conceptualisation would be a very different text from one that began with the first.

Positioning

The main participants in the testing process, we are told, are the test developer, the test user, and the test taker. Also often involved are the test sponsor, the test administrator and the test reviewer. Sometimes, many of these participants may be parts of the same organisation, with the notable exception, of course, of the test taker.

As clearly stated in Chapter 1, my position of value, my backdrop when I seek information about events, concerns the violations perpetrated on the participants in those events. So in the matter of testing, my focus is on the test taker, and in what ways the taking of tests and the inferences and consequences flowing from such events constitute a violation - a diminishing of personhood, a misrepresentation of potential or action, a claim to unwarranted accuracy of description, and thus unwarranted control and construction of the living human person who is taking the test.

The 1985 Standards acknowledge, with fine understatement, that "the interests of the various parties in the testing process are usually, but not always, congruent" (p1). This trivialisation of the traumatic effects, dislocations, and exclusions of millions of students based on test and examination results is quite remarkable. Perhaps it is just another example of the way social positioning can overwhelm interpersonal sensitivity and intellectual honesty.

The concern of the test makers and users is, after all, with hundreds, thousands, or hundreds of thousands of test takers (not to mention their concern with their Board of Directors and shareholders). But their concern is with them, viewed as a group. Their interest is with groups, not individuals; in summaries, not raw data; with simplifying complexities, not with complexifying individuals; with objectifying human subjects, not with subjectifying human events.

For the test constructor, sponsor and user there are so many difficult questions; so many criteria to consider; so many factors to consider if the overt and covert claims of the test makers are to be defended. We shall deal with these in due course. Yet to the test taker there is only one question, a normative question which emerges from his or her very construction as an individual. Have I passed or have I failed? Am I satisfactory or unsatisfactory? Am I normal or a nut case?

Additionally and ironically, it is precisely because they see the testing event from this individualised perspective, rather than from a group perspective, that they do not ask the more crucial, the more fundamental question: How much error, ambiguity, uncertainty, does this attribution contain? Or is it their powerlessness, and unheard voice, that makes these questions at the best unspeakable, at the worst unthinkable?

Sources of evidence

The 1985 Guidelines describes an ideal validation as including

> several types of evidence. . . Other things being equal, more sources of evidence are better than fewer. However, the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable validity (p9).

This is hardly reassuring for the test taker. The tautology and redundancy in the phrase "questionable validity" is remarkably inept; validity is proposed as the characteristic of the evidence used to support the construct "validity," and the essence of the concept is surely its very questionability. Far more damning, however, is the clear implication that evidence that does not cogently support the assertions of the test users should not be presented. Putting it another way, validity is a concept based on advocacy, is a rationalizating tool for a methodological decision already made, and is an ideological support rather than a scientific enterprise.

Is this an over-statement? Here is the first sentence of the next paragraph of the 1985 Standards: "Resources should be invested in obtaining a combination of evidence that optimally reflects the value of a test for an intended purpose" (p9). The word "optimally" says it all.

So, validity is clearly an advocacy construct, based on the assumption that any assessment data is innocent until proved guilty. The discourse about validity presents the case for the defence. There is no advocate for the prosecution, so the prosecution case does not present its case. More than this; the very idea of a prosecution case is denied by the definition of validity.

Yet here we also see, in the very heartland of post-positivist empiricism, the embryo of a discursive construct; an appeal, not to numbers, but to discourse. Over the next ten years Cronbach (1988) and Messick (1989a, 1989b,1994), doyens of psychometrics, in their born-again personas will enlarge the idea of construct validity to a point where Cherryholmes (1988) will nail it as fully discursive, and thus "linguistically, politically, economically, socially, culturally and professionally relative"(p450).

Even so, the advocacy position remains essentially unchanged. Messick(1989b) asserts that :

> To validate an interpretive inference is to ascertain the extent to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported. This represents the fundamental principle that both convergent and discrimanant evidence are required in test validation (p1).

But note the implication of "are less well supported" and its relationship to advocacy. And later in the same article, when he gets specific about invalidity implications of adverse social consequences, he says:

> If the adverse social consequences are empirically traceable to sources of test invalidity, . . . then the validity of test use is jeopardized. . . If the social consequences cannot be so traced - or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible - then the validity of the test use is not overturned (p11).

Note the use of the words "jeopardised," "less plausible," and "not overturned." Given the probabilistic nature of all social research, the chances of any test being declared invalid on the basis of these criteria, from this perspective, are remote.

Ultimately, Messick is eminently logical. For if "validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"(Messick, 1989a, p13), then even infinitesimal support, being support, makes the test valid, and nothing has really changed since Guilford's (1946) claim that "in a very real sense, a test is valid for anything with which it correlates " (p429). And as error will ensure that no tests correlate zero with anything, it follows that all tests are valid.

Reliability

Even though validity has taken on a post-modernist hue of recent times, reliability has, until recently, remained untouched as a "foundational" cornerstone of educational measurement. Reliability was seen as the lower limit of validity. An assessment could not be more valid than it was reliable.

The assessment industry, whether local, corporate, government, or quango, has embraced the reliability concept both ideologically and empirically. In contrast to validity, estimates of reliability are often obtained and circulated. There are two reasons for this: the reliability of the test can be measured using only data from the test scores; and often relatively high values (correlations of 0.7 - 0.9) can be obtained, if for no other reason that they are so constructed to ensure that such high internal consistency occurs.

Politically such reliability data can be used to "prove" the quality of the test, and maintain the illusion that reliability refers to "the degree to which test scores are free from errors of measurement," which is how they are described in the first sentence about Reliability in the 1985 Standards. In fact, the Standards emphatically insist that:

> For each total score, sub score, or combination of scores that is reported, estimates of relevant reliabilities and standard errors of measurement should be provided in adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test (p20).

Note that it is never suggested that the standard errors of measurement information should be available to test takers. There is a later chapter in the 1985 Standards entitled "Protecting the rights of Test Takers." Again there is not the vaguest suggestion here that such information should be made available to them.

However, even reliability is now under threat. Is there nothing sacred? Moss (1994), has cogently argued that there can be validity without reliability. She points out that:

> Reliability, as it is typically defined and operationalized . . . privileges standardised forms of assessment. By considering

> hermeneutic alternatives for serving the important epistemological and ethical purposes that reliability serves, we expand the range of viable high-stakes assessment practices to include those that honor the purposes that students bring to their work and the contextualized judgments of teachers (p5).

Such idiosyncratic behaviours and judgments tend towards a diversity that reliability abhors. There are two issues here. The first relates to the relationship between reliability and validity perceived from the standpoint of the assessors; the second deals with the concept of reliability, that is consistency, of performance as actually produced by the persons being assessed. The two issues are related in that they both relate to responses to persons involved in an event designed to describe what a person can do by asking them to do something else, and then making inferences about what they might do in another time and place and context.

Let's first look at this expectation of high reliability, and the theorising that precedes it. The argument is essentially this - if one test or examination is reliable then another similar test or examination will give the same verdict, however that verdict is communicated - as marks, grades, pass-fail, selected, or whatever. It is logical to assume, therefore, that one half of the test would give the same verdict as the other half, because all of the bits of the test contribute to the final score and hence the final verdict; putting it another way, we are dealing with some linear dimension here, some unitary idea or construct; all of the questions measure it with considerable error, but the more interconnected questions we ask, and the more inter-correlated answers we get, the more the error is reduced, and the more the measurement is refined to approach the true measure of it. Of one thing we are sure. The "it" is out there, waiting to be measured And "it" has a true value, that we can approach but never completely determine. This simplistic positivism is at the epistemological and ontological heart of educational measurement.

Teachers and public examination boards do not believe that this is what they are doing, even though the latter have no hesitation is using measurement theory to manipulate their results and rationalise their processes. They do not necessarily believe there is some unilateral trait or ability or skill that underlies the total score or grade. Indeed, as Willmott and Nuttall (1975) point out:

> In the field of 16+ examining it is quite possible that any increase in reliability would be to the detriment of validity. This is easily seen to be the case, since by refining questions and components so that they correlate highly is to learn more and more about less and less: the trait being measured is defined even more narrowly as reliability (in the sense of internal consistency) is increased. In such a situation, the validity of the examination concerned is bound to decrease owing to the narrowness of the field covered. A glance at any subject syllabus published by a CSE or GCE board shows clearly that the comprehension of a very wide variety of content is required of candidates and, in many cases, the educational objectives required of

> candidates in following the course are equally varied (p55).

It is a pity that these authors do not take this argument to its logical conclusion: that there is no single trait to be measured, that there is no linear concept to be categorised, and that there is no necessary correlation - indeed there may be some negative correlations, between the relative performances of candidates on various objectives. But this conclusion would lead inevitably to the final one, that there can be no meaningful rank order of students, because the rank order can give no meaningful information about the performance of individual students in relation to any particular objective (See Chapter 10 for a far more detailed description of the comparability issues involved).

Perhaps one more very simple example of this may be pertinent. Imagine a course in electrical wiring which has only two objectives; one relates to the safety requirements, the other to the ability to problem solve in practical situations. An examination is devised to measure the attainment on the course; half of the marks in the examination relate to safety requirements, and half to problem solving. Two students each obtain fifty per cent of the marks. What do we know about their attainment of the objectives? Nothing! One student may have got all the safety questions correct, and the other all the problem solving questions correct. In this case between them they may be considered to know everything, or nothing! In regard to validity, to inferences about objectives made from test scores, the validity has to be zero, if we focus on these individual students.

Note that the above argument is valid regardless of the correlations between the scores on the two parts of the paper for a group of students.

It can be seen that the reliability of the test in this case is irrelevant, as is any estimate of inferences that may be made about the group of students. For the group we could indeed make inferences about the probability that they knew, on average, a certain proportion of the safety information, and could solve a certain proportion of the problems. But just as a total score loses all the information about individual questions, so does it lose in this case all the information about individual students.

Incidentally, correlations across different subjects are often also of the order of 0.8. That is the correlation between two tests of different subjects is about as high as the reliability of any one test. (quoted by Nuttall & Willmott, 1975, p48). Perhaps there is a linear trait after all, but unrelated to the apparent construct being measured. What might this construct be? Traditionalists would be in no doubt that it was a general ability that they would label intelligence. Yet we know that the correlations between examination scores and other sorts of measures (eg, job performance) are very low, of the order of 0.3. So a more direct and sustainable interpretation is that "it" is the ability to perform in the events constructed around examinations. Examinations measure examination ability!

The second issue is rarely mentioned in the literature, and it relates to individual consistency of performance. An example might be taken from cricket. Batsmen vary in the consistency of their performance. Consider two batsmen who each has an average of about thirty runs over a large number of innings. One may score very consistently between 20 and 40 runs. Another may score the odd century, but may often make less than 5 runs. Test theory cannot account for this. It defines 30 as an approximation to their "true score," the score that best matches their "batting ability." But any deviation in a particular innings would be attributed to "random error," and be expected to assume a random rather than a consistent pattern. What becomes obvious from this example is that the average (true) score for these two batsman has a very different meaning; while for one it may indeed indicate the "most likely" score, for the other is indicates a most unlikely score indeed.

A fundamental contradiction

Now this argument, if we take it a little further, leads to a very strange conclusion. Let's go back to the first line of the Willmott and Nuttall (1975) quote: "it is quite possible that any increase in the reliability would be to the detriment of validity"(p55). They show why this is so in the measurement of any multi-dimensional area, and Moss (1994) indicates why it is so for "hermeneutical alternatives." But increase in reliability from what point? From 0.8, or from 0.5 ? Or from zero? Is there an argument to be made that all reliability negates validity. This would lead us to the apparently absurd conclusion that the greater the reliability the lower the validity, and the ultimately maximum validity is to be obtained from zero reliability. In terms of measurement, this would mean, of course, that human "constructs" were essentially unmeasurable. We can talk about them, but we can't measure them. Which is what Cherryholmes (1994) is really saying when he says the "construct validity is fully discursive." Isn't he?

In the next chapter I list thirteen sources of error, thirteen sources of invalidity. Two of these, related to multi-dimensionality and values, are dealt with by Willmott and Nuttall, and by Moss. What of some of the others? Do they show the same pattern of an increase in reliability leading to a decrease in validity?

Temporal errors are certainly increased by calculating reliability on the basis of one test at one time. As performance would be expected to vary with occasion and over time, one shot assessment certainly decreases validity error as it increases reliability

Contextual errors are certainly increased by confining assessment to pencil and paper situations and producing a very singular and artificial environment in which the assessment occurs, to the extent of standardising format and time available to complete the tasks. Again reliability is obtained at the expense of validity, which implies generalising to other contexts.

Construct errors are likewise increased through the limitations of content, form, process and media that is determined and narrowed through the testing or examination procedures. Again the capacity to generalise, and thus the validity, is diminished by the psychometric strictures required for high reliability.

The effect of high reliability on categorisation errors is complex. Where categorisation is defined in terms of percentiles of the group tested, categorisation errors are reduced as reliability increases, leading to an increase in validity. However, when one particular marking scheme (rather than another marking scheme) is used to increase the reliability, the reduction in categorisation error is illusory rather than real. And where comparability issues intrude, meaning fogs up as psychometric solutions compound the categorisation problems. So in these areas the effects of reliability on validity are moot.

In similar vein, errors attributable to frame of reference shifts, to labelling and attachment confusions, to prediction inaccuracies, or to logical type confusions, are largely indifferent to reliability. And whilst consequential errors, the negative effects of testing, have certainly been exacerbated by the quest for higher reliability, it is the quest rather than the empirical value that is involved.

Instrumental errors of course are reduced as reliability increases; indeed, reliability may be defined as the inverse of instrument error. So in this one area it is clear that increases in validity are dependent on increases on reliability. Yet if, as we have shown, the effect elsewhere is that such increase in reliability either decreases validity or has an indeterminate effect on it, then the general proposition holds, and we may say that in the empirical world, the procedures used to increase reliability result in a decrease in validity.

Born again validity

Messick (1989a) has broadened the concept of validity to refer to "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"(p13), and this includes the way values "influence in more subtle and insidious ways the meanings and implications attributed to test scores"(p59), so that "test validation embraces all of the experimental, statistical and philosophical means by which hypotheses and scientific theories are evaluated" (p14).

Messick's position seems to be generally accepted. The sources of potential error actually referred to do cover the range and depth of epistemological, ontological, and value sources referred to in this thesis. Yet even with this multiplicity of error, this proliferation of possibility of miscategorization, Messick (1989) insists that validity is a unitary concept, a singular "degree of support":

> The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force behind this integration is the trustworthiness of empirically grounded score interpretation, that is,

> construct validity(p5).

In other words, validity is a statement of faith in testing, a statement of justification by an "expert" that the whole assessment event is legitimate, is valid. Even though, in practice, for real tests, the considerations and scientific inquiries that Messick advocates are rarely carried out.

Let's look at this in more detail; first it is apparent that appropriateness, meaningfulness, the usefullness are sometimes quite separable. Appropriateness applies very much to particular values. In my value system, any test which violates individual students is inappropriate. Yet it might be quite meaningful in that some inferences made from it can be understood and acted on by teachers and administrators, and it may be useful in that predictions made from it help selection processes. In another case a test of inverted neuroticism may be quite useful in predicting successful medical students, but may be considered inappropriate for that application. It's meaningfulness may be moot. Ultimately, of course, the very meanings of appropriate, meaningful and useful are deferred; they are partial synonyms for valid, the word they supposedly elucidate.

It becomes clear that the "unifying force" then is not created by the congruencies among appropriateness, meaningfulness and usefulness, but rather by the "trustworthiness" of the "interpretation." In other words, by the power that resides in the status of the "expert" who controls the discourse in which the judgement is embedded. And because the discourse of validity is in essence about all the ways in which the measurement cannot do all the things it claims to do, and explicitly about some of the ways it might be done better, an advocacy judgment would concentrate on some way or ways in which the test was better than it might have been had such improvements not been made. According to Messick, this is the unifying force that asserts, and thus proves, validity.

Specifically, my analysis of Messick's (1989a) definitive paper in the third edition of Educational measurement indicates that he makes reference to over fifty sources of potential invalidity; for indeed, how can he describe how a test may be valid without focussing on all of the ways in which it might not be valid. I have indicated some of these references, and their relation to the error sources that I specify, in the next chapter.

Finally, the very existence of validity is established, validity is indeed made manifest, through the denseness of the arguments used to refute such existence, together with the reassurance that the battle continues, and some gains have been made.

Let me be specific: The definition of the construct of validity does not exclude the notion of invalidity. However, the discourse on validity, constructed as it is from the position of advocacy, excludes the notion of invalidity as an issue. More than this, the discourse itself becomes the arbiter of the proof of validity claims, independently of empirical data, which becomes irrelevant within the density and complexity of the discourse; as a result, empirical data to justify validity claims is rarely collected, and when it is it is inevitably construed as supporting the claim. Evidence rejecting the validity claim is never collected because such positioning is absent from the discourse. Madaus (1986) puts it nicely:

> present methods of gathering content validity evidence are inadequate; they are designed in such a way as to almost guarantee a positive outcome. Alternative methods designed to disconfirm or test counter hypotheses about the issues are, in my experience, never employed (p12).

Practically, the psychometric scam is accomplished by focussing on the test score, and ignoring its dark side, the standard error of estimate; specifically, by implying that the estimated score is the true score, that the intention is the empirical fact, that talking about problems of validity magically increases validity, and that increasing validity makes a test valid.

Validity and the predominant paradigm

When advocacy is positioned, aligned to the predominant paradigm, then advocacy is interpreted as truth. Truth not as the production of true utterances, but in Foucault's (1982) sense of "the establishment of domains in which the practice of true and false can be made at once ordered and pertinent"(p8). From the 1980s, when the prevailing societal metaphor is the discourse that surrounds economic rationalism, and in particular those myths connected with people competencies, the metaphor is rabidly post-positivist, and validity definitions (advocacies) based on those assumptions will be seen as self-evidently true. As Cherryholmes (1988) puts it from his post-modern perspective: "boundaries limiting construct-validity discourse have yet to be justified. They are policed nonetheless "(p154).

In contradistinction, advocacies for more post-modern descriptions (eg validity characteristics for qualitative research) are clearly not aligned to the prevailing world-view, and so will be interpreted as justifications. They advocate from a loser's position, so at the best their views are accepted as tentative, at the worst as unproven and hence unacceptable assumptions. This is inevitable because no abstraction can be proven to be correct, so acceptance is always a function of value, rather than of rational proof; and moral value is usually construed as stabilisation of the status-quo, as confirmation of the predominant paradigm.

Shepard (1991) gives an example: "measurement specialists asserted that performance assessments are less reliable and less valid than traditional tests and that they are potentially biased because they rely on fewer tasks." But then she adds: "Why are existing tests presumed to have the high ground in this dispute? What claim do traditional tests have to validity?" (p10).

This is not to deny the acceptance of such advocacy in localised communities (eg some faculties of some Universities) where a paradigm shift has already occurred.

Qualitative assessment and qualitative research

Validity criteria in qualitative assessment has lagged behind validity in quantitative research. However, the two fields are closely aligned. In fact Messick (1989a) regards then as virtually synonymous in that

> test validation in essence is scientific inquiry into score meaning - nothing more, but also nothing less. All of the existing techniques of scientific inquiry, as well as those newly emerging, are fair game for developing convergent and discriminant arguments to buttress the construct interpretation of test scores" (p56).

I do not want to focus on the blatant advocacy aspects of this statement implicit in such terms as "fair game" and "buttress," but rather on its implication for using research validation criteria for assessment. In addition, I would want to include "categorisations" as a limiting aspect of "score." With this addition the work done on qualitative criteria for research validity becomes appropriate for assessment validity.

Summary

We have worked our way through some of the minefields of validity and reliability discourse. In particular I have indicated how the notion of advocacy built into the very definition of validity overwhelms scientific detachment, and effectively silences the logical inferences that derive from the voices of confusion and error that are the very basis of validity discourse.

The emphasis on reliability of assessment instruments is also shown to be a misplaced source of credibility for assessment, because measures to increase reliability are shown to decrease validity.

Now the coin can be flipped. The underside of validity can be examined. The nastiness of error can be exposed. In the next chapter the sources of invalidity are spelt out in detail.

Part 5: Synthesis

Chapter 17: Error and the reconceptualising of validity

Preview

From the analysis so far, it is possible to produce a general definition of error as it applies to the field of educational measurement and/or categorisation. This is the flip side of validity which exposes that general nastiness called invalidity.

In this chapter the notion of invalidity is reconceptualized, having both discursive and measurable components. Thirteen (overlapping) sources of error are examined, all contributing to the essential invalidity of categorisations of persons. For easy reference I have indicated the summary theoretical and practical definitions of these error sources in bold print.

Definition of error

Error is predicated on a notion of perfection; to allocate error is to imply what is without error; to know error it is necessary to determine what is true. And what is true is determined by what we define as true, theoretically by the assumptions of our epistemology, practically by the events and non-events, the discourses and silences, the world of surfaces and their interactions and interpretations; in short, the practices that permeate the field.

All assessment statements about a person are statements about that person engaged in an event, or a potential event. They are descriptions or indicators or inferences about the person's performance in that event. As such they involve at the very least an event in which the person being assessed is an element, and an event in which the assessor engages directly in the first event, or with a product (element) of it.

Error is the uncertainty dimension of the statement; error is the band within which chaos reigns, in which anything can happen. Error comprises all of those eventful circumstances which make the assessment statement less than perfectly precise, the measure less than perfectly accurate, the rank order less than perfectly stable, the standard and its measurement less than absolute, and the communication of its truth less than impeccable.

I want to list some of those sources of error, some of the conditions that change the measurement of a standard from a thin red line into a broad blue band: In doing so I will reject the notion of construct validity as a unitary concept, and dismember its dark side into disparate if sometimes overlapping categories.

Sources of error

I have named these sources of error:

1. Temporal errors
2. Contextual errors
3. Construction errors
4. Labelling errors
5. Attachment errors
6. Frame of reference errors
7. Instrument errors
8. Categorisation errors
9. Comparability errors
10. Prediction errors
11. Logical type errors
12. Value errors
13. Consequential errors

1. Temporal errors

We would hope our description of performance would have some substance; would be a stable quantity, invariant over time and space, rather than some ephemeral numerical butterfly attaching itself momentarily to the person assessed. If the person's performance is described differently if done at another time, in another place, with another group of people, then such difference as there is represents a source of error.

Or is it? Should we rather discount stability as being counterproductive in an educational situation? If stability is seen as the very antithesis of the educational enterprise, which we could define as being dedicated to change, then we would not wish any description to remain stable, as this would represent a nullification of the educational process.

Contrarily, if we wish to maintain stability as a criteria for assessment accuracy, we must be certain that all learning pertaining to the performance ceases at the time of assessment. And that none occurs during the assessment process. As well as all forgetting for that matter. Otherwise the error of the description increases rapidly, as the permanency of the description becomes increasingly dismembered by the ravages of time.

Regardless of which side of the fence we want to sit, or whether we want to sit on the fence, pretend it isn't there, and attribute the concomitant pain to other variables, stability must logically remain as a pertinent, or in conventional circles an impertinent criteria, to be considered in any estimate of error in assessment. My conclusion is that the logic of its contradictions makes most of the academic and psychometric definitions of reliability trivial.

So temporal errors have their genesis in changes that occur over time; persons change over time; tests change over time; the "same" event has different meanings over time. People are not computers, they react differently at different times; and they forget. So temporal errors increase over time. (Not to mention that different people make different meanings out of the same event; which makes it, of course, a different event.)

Temporal errors thus include all those confusions that constitute the dark side of stability, one aspect of reliability.

Practically, temporal errors are indicated by the differences in assessment description when the assessment occurs at different times

2. Contextual errors

Contextual errors constitute the underside of claims to generality and generalisability.

Any performance is relatively specific and defined: It is a single instance of possible instances; it is an event chosen from a multitude of possible events; it is a particular designed to illustrate a generality. Yet the performance will invariably be described (labelled) in terms of the generality it aspires to, rather than the specifics that define it. This is true of almost any evaluation, any test that goes beyond the description of a single behavioural objective, and even that, one step back, will often be found to be illustrative of a class of objectives, rather than of particular significance in its own right.

In the old days (good or bad depending on our values), this would constitute an example of "transfer of training." The claim was that if you could think clearly in Latin, then this should transfer to dealing adequately with the complexities of life in the social world; or if you could think logically in mathematics, then you could do so in international affairs; not to mention playing Rugby being a necessary prerequisite to running an Empire. When empirical data showed that such transfer was tenuous, the notion was kept, but the name changed. Taxonomic terms such as application and analysis, or the more up-market process called problem solving, have latterly laid claim to this temporarily non-habitable area. As well, the notion of a "skill" has latterly become fashionable, and generalisable social, cognitive, emotional, spiritual, and psychomotor skills proliferate, securely untrammelled by prophylactic empirical data of any kind.

As soon as assessment descriptions are committed to paper, their material permanency is dramatically increased. Likewise, the span of their associations is spread and emphasised. No longer just a description of a particular performance, the assessment becomes interpreted as a measure of knowledge and ability, an indicator of achievement on a course of study, and a predictor of future success or failure.

One source of error then is the magic transformation that occurs between numbers and categorisations, between specific acts and generalised descriptions. Unless the assessment statement purports to be no more than a statement about a particular assessment event, then the differences between this statement, and those obtained from all other possible contexts, is error; these are the generality differences attributable to other equally relevant contexts, eg written, oral, cooperative, on-the-job; all those boundaries that possibly could contain the assessment event that are different to the boundaries of the particular assessment event. Context also includes those power relations that pervade it and the judgment processes embedded in it that affect the performance of the person assessed, and the judgment of the person assessing; and this includes those that the boundary localises, as well as those that invade its permeable surface.

Contextual errors contain all the ambiguities inherent in those relations and elements and discourses that impinge on the event, but get excluded from the label.

Practically, contextual errors include all those differences in performance and its assessment that occur when the context of the assessment event changes.

3. Construction errors

The performance that is described in an assessment is generally built up of a number of parts; a science test is built up from a number of questions; an electrical automotive practical test requires the identification and repair of a selection of common electrical faults; a social skills assessment requires gradings on a number of interactional criteria, or more likely a game constructed about such criteria in multiple choice form. Such constructions are designed to represent the course of study, or the skill requirements, or the criterion referenced framework, that the assessment is supposed to describe. Further back still, the course has itself been constructed to improve performance in some areas of living, in some role as citizen, home maker, academic, engineer, baker, or whatever.

Somewhere, sometime, someone must make a choice about how far back along the chain of constructions we go in order to estimate the error, the difference between the "perfect" description of performance and the actual one that our assessment produces.

Let's take the electrical automotive test as an example. We could begin with a requirement to describe how well a student could identify and repair any electrical fault on any car brought into any garage (A). From this we construct a thirty hour course of study called Automotive Electrical Mechanics 2M, complete with course aims and objectives and assessment criteria (B). From this we construct a one hour pencil and paper test (C) and a two hour practical assessment (D).

Now how are we to describe the construction error in assessing a particular person? Is it the difference between the descriptions given in C and D? Or the difference between the matches of B and C on the one hand, and B and D on the other? Or should we look at the matching between C and D and A? Or is it all of these?

> *Why don't you describe A directly?*
>
> *You mean put people who've done the course into a garage and see how they perform?*
>
> *Yeah. Why don't you do that?*

> *It would be very expensive to do it for everyone?*
>
> *You don't have to do it for everyone. Just for enough people so you'd know if there was an error.*
>
> *There's always going to be an error.*
>
> *OK, so you find there's an error. If it was a small one then you could assume that the course, or the test at the end of the course, was well constructed because it did what it was supposed to do.*
>
> *That would be nice.*
>
> *And if there was a big discrepancy then you'd have to do it different.*
>
> *Do what different?*
>
> *I dunno. You're supposed to be the expert. Do the end-test different. Or do the course different. Or it might be easier to find another garage.*
>
> *Bit dangerous. There could be a lot of people get upset if we did that. No telling what sort of litigation we might run into if we found that the course didn't do what we said it would do.*
>
> *So ignorance is bliss, huh?*
>
> *Certainly not. We just need to be very careful, in terms of spending time and money on obtaining information that at the best will be useless, and at worst will only erode confidence and create instability.*
>
> *Like I said. Mum's the word!*

So error is immanent not only in the selection that determines the content and process of the assessment event, but also in the choice about what aspects will be elucidated in the assessment description.

Construction errors contain all of those errors in sampling, all the idiosyncrasies and biases that are contained in the construction of a specific test or set of demands that constitutes one element of the assessment event: these include not only the construction of the test content, of its elements, but also the construction of its form and style. Construction errors include all those generality errors attributable to the performance task itself, rather than to its timing or its context.

Practically, construction errors are indicated by all those differences in assessment description when the same construct is assessed independently by different people in different ways.

4. Labelling errors

Assessing is about describing some human performance. To give it a meaning the "some" must be specified: performance in typing; skill in mathematical problem solving; a dramatic presentation. So regardless of frame, it is necessary to specify in some way what it is that is being described. We must label the area of performance in some way, for otherwise it cannot be communicated.

The meaning of a communication is its reception, not its intention. In assessment the label is the message which is intended to describe a particular area of performance - involving particular knowledge, understandings, skills, processes, or whatever. The label has a particular meaning for the assessor growing out of this intention. Different meanings before the event will result in different assessment events being constructed to fit the label.

What meaning the assessed, or any other person who has access to the label, gives to it, is moot. But of one thing we may be certain. The meaning will not be identical to the meaning intended. The difference may be slight, or immense, but regardless of the magnitude will represent, at a fundamental level, an error (Korzybski 1933). Different meanings after the event will result in different interpretations of the assessment label, different inferences about what it implies.

An assessment must be an indicator of something. It must have a name. Differences in the meaning of the name, both before and after the event, constitute confusion and hence error. Labelling errors are defined by all the differences given to the meaning of the assessment (what it actually measures) by all the participants in the assessment event(s), and by the users of the assessment information.

Practically, labelling errors are indicated by the range of meanings given to the label by all those who use it before, during or after the assessment event.

5. Attachment errors

There is a further issue in regard to labelling. Once the label has been marked in some way, once the description is attached to it, where is it pinned? Does it belong to the person assessed? Is it more a description of the assessor? Does it represent some quantity or quality that might more appropriately hover somewhere in the space between, a relational field vector describing a complex interactional phenomena involving task, performance, assessor and assessed?

Given my ontological stance that all information is information about events, it follows that any attempt to attribute such information to a particular element of the event involves a fundamental epistemological error. To the extent that all other elements and conditions are held constant and overtly included in the description, to that extent is the simplification of language involved in the specific attribution partially justified; but such specificity of the conditions of the event tends at the same time to increase contextual error.

Attachment errors are the ontological slides that occur when a description of a relational event is attached to one of the elements of that event; specifically, when a complex relational event involving the construction of a test, an interaction of the test with a person, and a judgment of an assessor, is described as a property of the assessed person, this is an error in attachment.

Practically, attachment errors are indicated by the specification of those elements and boundaries of the assessment event that have become lost in the assessment description.

6. Frame of reference errors

Within the assessment arena are four competing definitions of the true, the correct, the impeccable. It follows that there are four associated notions of error. To the extent that the definitions of assessment truth, or more specifically the assumptions underlying them, are contradictory, so will be our methods for reducing error in the different frames; further, to the extent that the frames are confused, to that extent is error compounded. (See Chapter 13).

Frame of reference errors are defined by all those confusions and category differences that occur because of the different stable assumptions of the four frames of reference for assessment, as well as those contradictions and confusions that occur when shifts occur between frames during the assessment process.

Practically, frame of reference errors are indicated by specifying the frame in which the assessment is supposedly based, and indicating any slides or confusions that occur during the assessment events.

7. Instrument errors

Any measurement requires a measuring instrument. So any rank ordering, grading or scoring involves some measuring instrument; at the very least, such an instrument must attend to questions of calibration, which involves scale, replicability, and theory-practice bridging. Any claims to measurement must relate to some defined Standard scale. Whether the instrument is a test of some sort, or is assumed to have some material reality inside the mind of an examiner, all measuring instruments contain errors in mechanisms and hence in their readings. (See Chapter 9)

When psychometric theories are used, instrument errors are fed by all of the discrepancies between the theory and the empirical data, and are intrinsic in all of the notions of probability that pervade such theories.

Instrument errors then contain all those uncertainties of calibration, all those anomalies of replicability, all those confusions and discrepancies and mis-matches in theory-practice bridging, that are involved in the determination of the rank order, in the making of the mark, in the determination of the measure.

Practically, many aspects of instrument error are covered by other category errors. To avoid unnecessary overlap, I will limit the practical indicator of instrumental error to those errors implicit in the construction of the measuring instrument itself; what is conventionally called standard error of the estimate.

8. Categorisation errors

Any categorisation involves a comparison between a standard of acceptability, and a particular measurement or judgment about adequacy or quality.

Categorisation errors derive from confusions about the definition of standard of acceptability, from differences in the meaning of what is being assessed and in the magnitude of its measurement, and in the variability of the judgment process in which the comparison with the standard is made. (See Chapter 11)

Practically, categorisation errors are all those differences in assessment description that occur when particular data is compared with a particular standard to produce a categorisation of the assessed person.

9. Comparability errors

Comparability errors occur whenever assessment scores are added to produce a total score. Public examinations and grade point averages are examples of such summations, as are any qualitative assessments involving more than one criteria. What such additions mean, and who is privileged by such additions, are questions inherent in the process.

Comparability errors include all those confusions about meaning and privileging that inhabit the addition of test scores, grades or criteria related statements.

Practically, comparability errors are indicated by constructing different aggregates according to the competing models. The differences that these produce indicate the comparability error.

10. Prediction errors

Implicit in most assessment, and explicit in some, is the notion of prediction. Whilst the idea of generality contains some element of logic in its derivation, prediction can be pure magic - correlation without connection is very possible, and is not predicated on causal relationship. It has been reported that the number of storks sighted over London is correlated with the number of births in that city, and thus may be used as a predictor. The causal relation here is moot.

More seriously, many assessment descriptions are overtly or covertly connected to expectations about future performance. High school grades are presumed to be related to success at College or University. School performance is expected to relate to job success. Trade courses are designed to improve quality of performance in the workplace. So assessments on those courses might be expected to correlate with later performance. Yet even if they do, this in no way proves there is any causal link.

The criterion measures themselves are often problematic; most practical criterion measures themselves involve an assessment, subject to all of the sources of invalidity and error that dogged the original assessment. High predictive correlations may occur because both assessments are measuring something other that what they are described as measuring; for example, the ability to perform in competitive, written events, independent of the content. And low predictive correlations may mask genuine positive relationships because of all the errors entailed in the assessments, though such "genuine relationships" must forever be hidden, relegated to fantasy because divorced from empirical sustenance. Alternatively low correlations may mask the reality of relative homogeneity of performance status, or of genuine multi-dimensionality of that performance.

So interpreting the meaning of high correlations can be quite tricky. For example, if the rank order of students on a university entry examination in Physics correlates 0.9 with their first year Physics results at University this could be interpreted as an enormously successful outcome in terms of educational prediction. It is also completely consistent with the implication that no new Physics has been learnt, or that the University course has been completely unsuccessful in compensating for initial inequities in knowledge and opportunity.

What becomes apparent is that this area of prediction, which on the surface seems very amenable to empirical verification, is fraught with errors of interpretation which are neither measurable nor resolvable. Positioning and power relations will largely determine the trend of the discourse, and whether such discourse becomes a verification of validity, or an explication of error.

Explicit or implicit in most assessments is the claim that they relate to some future performance, that they predict a particular product from some future event, a quality of some future action. Prediction error is the extent to which these predictions, and the subsequent events, are not identical.

Practically, prediction error is indicated by the differences between what is predicted by the assessment data, and what is later assessed as the case in the predicted event.

11. Logical type errors

Test scores are often interpreted as giving specific information about what a student can or cannot do. For example, a score of 90 per cent on a spelling test gives no information about whether any individual item on the test was actually spelt correctly by a particular student. Any assumption to the contrary is a logical type error. Similarly, a score of 80 per cent on a mastery test gives no information about what information or skill has been mastered. Common inferences made from test scores are riddled with such logical type errors.

Logical type errors occur whenever there is confusion between statements about a class of events, and statements about individual items of that class.

Practically, logical type errors are made explicit when the explicit and implicit truth claims of a particular assessment are examined and any logical type errors are made explicit. Such exposure may invalidate such claims.

12. Value errors

All tests and examinations involve the construction of questions and the interpretation and valuation of answers. As such they are explicit and implicit statements about value; these particular questions, and these favoured answers, are implicit statements about what knowledge, actions, processes and interpretations are valued. And by implication, which are not so valued. Such implications move well beyond content; style and form and medium are of equal or more importance.

To the extent that the values implicit in the assessment event are not explicit, are contested, or are contradictory, to that extent is the assessment event invalid with respect to value. To the extent that the assessment event(s) and the event about which inferences are made are incongruent in terms of their value assumptions and emphases, to that extent is the error component engorged.

Practically, value errors are indicated by making explicit the value positions explicit or implicit in the various phases of the assessment event, including its consequences, and specifying any contradiction or confusion (difference) that is evident.

13. Consequential errors

Messick (1989a) and Cronbach (1988) both accept that the effects of testing have to be taken into account when assessing the validity of testing. It follows that any distortion of learning through the assessment process constitutes a source of error.

To take this view, however, is to make an extension to the meaning of validity, or of invalidity. For we have to ask, in what way does such distortion of learning detract from the appropriateness, usefulness, or meaningfulness of the inferences made from the test scores? Are the test scores less useful because they have distorted the learning process? Certainly in such a situation the testing process has been counterproductive, which is a good reason for dismantling it, if learning is a major purpose of education. However, earlier chapters have shown this to be a naive proposition. Assessment has other more important if less salubrious social purposes.

Logically, distortion of learning increases error only if we take error to include not only the differences between what the test measures and what is or might be, but also between what the test measures and what might have been. This seems to take us into a rather transmogrified realm. Even so, any distortion of learning possibilities contributes to the violation of those persons whose learning, and possibility of growth, is thus diminished. And as that very learning is part of the event that the assessment presumes to measure, then it is legitimately included as inappropriate, and thus a source of error, a (retrospective) interactive interference effect.

Consequential errors involve all those negative effects on a student's learning and a teacher's teaching that are attributable to the assessment event. (To the extent that it produces inequity among sub-groups, positive effects on learning may also be involved).

Practically, at a simplistic level, consequential errors are indicated by the differential positive and negative effects that individual teachers and students attribute to the assessment process: At a more profound level it involves an explication of the very construction of their individuality, and all of the potentially violating consequences of those constructions. (See Chapters 4 & 5)

Invalidity according to Messick

Messick's (1989a) treatment of Validity in *Educational Measurement* is an excellent review of current (theoretical) state of the art, progressive in stance, and its implications vastly surpass current practice.

In this section Messick's work is looked at from the standpoint of invalidity, in order to indicate that the sources of invalidity indicated above are indeed well-established, if somewhat opaquely discerned, in the literature on validity.

Temporal error

Here are two passages from Messick that illustrate some of the temporal problems of validity. The first relates to the lack of necessary conjunction between construct meaning on the one hand, and stability of measure on the other:

> In regard to temporal generalizability, two aspects need to be distinguished: one for cross-sectional comparability of construct meaning across historical periods . . and the other for longitudinal continuity in construct meaning across age or developmental level. It

> should be noted that individual differences in test scores can correlate highly from one time to another (stability) whether the measure reflects the same construct on both occasions (continuity) or not. Similarly, scores can correlate negligibly from one time to another (instability), again regardless of whether the measure reflects the same or a different construct (discontinuity) on the two occasions (p57).

So even if the measure remains the same at different times, it may mean different things. And if the measure is different at different times, it may mean the same!

Here is the second example. Messick argues that it is not necessary to assume that

> the more generalizable a measure is, the more valid. This is not generally the case, however, as in the measurement of such constructs as mood, which fluctuates over time; or concrete operational thought, which typifies a typical developmental stage (p57).

From the standpoint of invalidity, that a test is invalid unless proved otherwise, how could the measurement of such an ephemeral quality ever be validated?

Contextual error

Contextual errors receive a lot of attention from Messick. Here is one example:

> Tests do not have reliabilities and validities, only test responses do. . . . test responses are a function not only of items, tasks, or stimulus conditions but of the persons responding and the context of measurement. This context includes factors in the environmental background as well as the assessment setting. . . . Thus, the extent to which a measure displays the same properties and patterns of relationships in different population groups and under different ecological conditions becomes a pervasive and perennial empirical question (p14-15).

This certainly captures the idea that the data belongs to a complex event, even though Messick does not follow through to the logical conclusion that the test score data cannot then be detached from the event and attached to an individual.

Moreover, in terms of error in individual measures he misses the point; for even with knowledge of the relationships between test measure - group - context, we still have no knowledge about the specific error in an individual score. (In group terms it could be anywhere within plus or minus three standard errors from the estimate).

Here is another example that raises the more fundamental issue of context as boundary condition:

> studies of the transportability of measures and findings from one context to another should focus on identifying all of the boundary variables that are a source of critical differences between the two contexts, as well as gauging the potency and direction of the effects of these boundary variables on events in the two conditions (p58).

Indeed, for science is nothing if it cannot adequately define the boundary conditions within which the limited experimental events that define its world can be controlled. So the assessment is invalid unless all the boundary conditions (that cause unexplained variance) can be specified. And, of course, they never can be.

Construction errors

Construction problems are often dealt with in terms of content validity. Messick comments that "the heart of the notion of so-called content validity is that the test items are samples of a behavioural domain or item universe about which inferences are to be drawn" (p36). He has some problems with this, for "to achieve representativeness . . . one must specify not only the domain boundaries but also the logical psychological subdivisions or facets of the behaviour or trait domain" (p39). Furthermore, "in point of fact, items are constructed, not sampled" (p40). And finally, Messick's crunch point :

> knowing that the test is an item sample from a circumscribed item universe merely tells us, at the most, that the test measures whatever the universe measures, and we have no evidence about what that might be, other than a rule for generating items of a particular type (p40).

So even the apparently simple task of getting some test questions together is fraught with difficulties, again justifying a invalidity label until compelling evidence is presented that these problems have been solved.

Labelling errors

Messick is adamant that "the meaning of the measure . . . must always be pursued - not only to support test interpretation but also to justify test use" (p17).

At least some of this meaning is carried by the construct label, and "constructs are broader conceptual categories than are test behaviours, and they carry with them into score interpretation . . . the evaluative overtones of the construct labels (p59).

One such problem with the label is how broad to make it. Messick spells out the dilemma:

> In choosing the appropriate breadth or level of generality for a construct and its label, one is buffeted by opposing counterpressures toward oversimplification on the one hand and overgeneralization on the other. . . . choices on this side (of oversimplification) sacrifice interpretative power and range of applicability as the construct might be defensibly viewed more broadly. At the other extreme is the apparent richness of high-level inferential labels such as intelligence, creativity, or introversion. Choices on this side suffer from the mischievous value consequences of untrammelled surplus meaning (p60).

Another problem with a label that applies to everybody is that different people do things in different ways:

> In numerous applications of these various techniques for studying process, it became clear that different individuals performed the same task in different ways and that even the same individual might perform in a different manner across items or on different occasions. . . that is, individuals differ consistently in their strategies and styles of task performance. . . this has consequences for the nature and sequence of processes involved in item responses and, hence, for the constructs implicated in test scores. . . test scores may mean different things for different people. . . for different individuals as a function of personal styles and intentions. . . Indeed, . . . a test's construct interpretation might need to vary from one type of person to another (p54-5).

So why not from one person to another? In this regard note that validity has always been a group concept. Human rights, with its associated absence of violence, is a term that applies to individuals and not to groups; to claim that 95 per cent of a population is not subjected to human rights violations such as torture, incarceration and extermination is hardly a claim for a good human rights record. Why is assessment any different?

It would seem from Messick's own example that the label must be individualised in meaning before it can validly be applied to an individual person.

Attachment errors

The idea that assessment data gives information about an event rather than about a person is contrary to the very conception of assessment in general, and to psychometrics in particular. However, there are glimmerings of light in Messick's work that are encouraging. Here are two examples:

> The possibility of context effects makes it clear that what is to validated is an interpretation of data arising from a specified

> procedure (p15).
>
> . . . the important validity principle embodied by this term (trait validity) might be mistakenly limited to the measurement of personal attributes when it applies as well to the measurement of object, situation and group characteristics (p15).

In the first quote the data is seen to be related to a procedure, that is, an event involving relationships; in the second the validity, if not the data, is seen clearly not to be limited to the personal.

Frame of reference errors

Messick does not mention frame of reference errors in the form that I have developed them in this dissertation. However he does talk of the various theoretical frameworks for intelligence, including the two well-known "geographic" models of intelligence as a single dimension, or as multiple discrete abilities. And then goes on to mention a computer model, an anthropological model, a sociological model and a political model. He then comments:

> If two intelligence theories sharing a common metaphorical perspective - such as uni-dimensional and multi-dimensional conceptions within the so-called geographical model - can engender the different world phenomenon of investigators talking past one another, as we have seen, just imagine the potential babble when more disparate models are juxtaposed (p61).

A close inspection of the literature on assessment obviates the necessity to imagine, for fact is indeed stranger than fiction, and indicates the massive sources of invalidity from this source.

Instrument errors

Instrument errors as such don't get much attention in this work, perhaps because, as I have defined them, they are an aspect of reliability rather than of validity, and so are dealt with in a different chapter in Educational measurement (Linn, 1989a).

However, he does note that "the very fact that one set of behaviours occurs in a test situation and the other outside the test situation introduces an instrument error"(p37), indicating that he is aware of a fundamental shift in context that pervades the use of tests for assessment.

Categorisation errors

About the validity of any particular categorisation Messick is remarkably silent. A short section on decision models of cost - benefits is all that scratches the surface of the chasm of silence (p78-80). This despite the fact that in practice the meaning of the categorisation assumes more importance than the meaning of the construct; to the individual student the distinction, or the failure, is more important than whether the assessment measured what it claimed to measure.

The substantiality of the standard is a necessary prerequisite to the allocation of a measure to a category. Or, for that matter, of the conversion of a category to a measure, as in a conversion of "better or worse" to "more or less." Are standards then irrelevant to construct validity, which in Messick's model is all validity? For surely the construct meaning given to a test score is submerged in the social world, in most cases, under the weight of its categorisation as a grade. To limit the definition of validity to test scores hardly affects the issue, because surely the categorisation then becomes the first interpretation, the first utility, the first action, and hence a crucial element in the validity discourse.

Should I really have been so surprised, as I most genuinely was when I realised for the first time, as I wrote the two preceding paragraphs, what had occurred? Was it a conscious decision on Messick's part not to include the categorisation issue in his extremely comprehensive study? Or is the erosion of the problem of the standard from professional and public memory so complete. Certain it is that though I have been very familiar with Messick's chapter for four years, and standards are my major area of interest, I had not noticed the almost complete omission of any treatment of the issue in his definitive paper on validity till now.

Whatever, categorisation errors remain a major source of invalidity in assessment, and without clear evidence to the contrary, must be assumed to be very large indeed, making most categorisations of individuals invalid.

Comparability errors

Now whilst Messick is certainly aware that "a single total score usually implies a unitary construct and vice versa" (p44), he does not develop many validity implications of this until he begins to discuss test-criterion relationships. He makes the point that "criterion measures must be evaluated like all measures in terms of their construct validity"(p70). He seems to accept that most criterion measures are "multiple and complex." He points out that it does not "make such sense logically to combine several relatively independent criterion measures . . . into a single composite as if they were all measuring different aspects of the same unitary phenomenon"(p74). He goes on to state that:

> On the contrary, the empirical multidimensionality of criterion measures indicates that success is not unitary for different persons on the same job or in the same educational program or, indeed, for the same person in different aspects of a job or program. furthermore, because two persons might achieve the same overall performance levels by different strategies or behavioural routes, it would seem logical to evaluate both treatments and individual differences in terms of multiple measures (p74-5).

Easy to say, of course, but much harder to do. Because this leads inevitably to the use of "judgmental weights that reflect the goals or values of the decision maker"(p75), which leads directly into all the confusions and errors dealt with in the chapter on comparability.

Prediction errors

Messick discusses prediction errors under the general rubric of test-criterion relations and decision making (p69 -88). He points out that "the major threats to criterion measurement . . . are basically the same as the threats to construct validity in general" (p73). In other words, errors are compounded in prediction errors because the errors in the test are multiplied by the errors in the criterion measure. In addition "other biasing factors include inequality of scale units on the criterion measure, which is a continual concern when ratings serve as criteria, and distortion due to improperly combining criterion elements into a composite" (p73). He talks of "inappropriate weights . . . applied to various elements in forming composites" (p73), yet who could say what an "appropriate" weight was?

So one source of confusion is whether the criterion domain "entails a single criterion or multiple criteria" (p74). He concludes that:

> use of measures of multiple criterion dimensions or components affords a workable approach to composite criterion prediction . . . by combining correlations between tests and separate criterion dimensions using judgmental weights that reflect the goals or values of the decision maker (p75).

Maybe, but this takes us into further sources of confusion related to differing values, differing goals, of different decision makers, and a concomitant further proliferation of error.

Value errors

In terms of the validity of tests, Messick is adamant that "the issue is no longer whether to take values into account, but how" (p58). It follows that "because validity and values go hand in hand, the value implications of score interpretation should be explicitly addressed as part of the validation process itself" (p59).

He is also clear that "data and values are intertwined in the concept of interpretation"(p16), and furthermore, "values . . . influence in more subtle and insidious ways the meanings and implications attributed to test scores with consequences not only for individuals but for institutions and society" (p59). So it is not only obvious biases expressed in interpretations that we are dealing with here, but "more subtle" mechanisms.

For example, not only are "some traits . . . open to conflicting value interpretations" (p60), (shouldn't this read "all traits"), but "the tenability of cause-effect implications is central, even if often tacitly, to the construct validation of a variety of educational and psychological measures such as those interpreted in terms of ability, intelligence, and motivation" (p58). So if cause-effect thinking is shown to be simplistic and epistemologically bankrupt in a more ecological world-view, where does that leave such "traits"?

So Messick centred his attention

> on the value implications of test names, construct labels, theories and ideologies, as well as on the need to take responsibility for these value implications in test interpretations. That is, the value implications, no less than the substantive or trait implications, of score-based inferences need to be supported empirically and justified rationally (p63).

Here Messick makes a brilliant case for the fundamental invalidity of all test data on the basis of value confusion and hence inability to interpret meaningfully test measures.

Consequential errors

Messick pays considerable attention to the consequential basis of test validity (p58-63). By this he means "the often subtle systematic effects of recurrent or regularised testing on institutional or societal functioning" (p18). He is firm that "social consequences cannot be ignored in considerations of validity" (p19). He then spells it out in more detail:

> The consequential basis of test interpretation is the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the ideologies in which the theory is embedded. A central issue is whether or not the theoretical implications and the value implications of the test interpretation are commensurate (p20).

This may well be a central issue, but surely not the central issue. They may be commensurate and yet be utterly unequable to groups or to individuals. Messick himself acknowledges this later when discussing cost-benefit decision making:

> This concern with minimizing overpredictions, or the proportion of accepted individuals who prove unsatisfactory, is consistent with the traditional institutional values of efficiency in educational and personnel selection. But concern with minimizing underpredictions, or the proportion of rejected individuals who would succeed if given the opportunity is also an important social value in connection both with individual equity and with parity for minority and disadvantaged groups (p80).

Exactly, and Messick is equally precise when on the next page he concludes that "in practice, however, such balancing of needs and values comes down to a political resolution" (p81). That is, a solution based on power relations, which are inevitably asymmetrical. So if we are to be clear about invalidity errors of a consequential nature, we had best be mindful of the mechanisms through which such power relations are distributed and applied.

Messick's fudged solution

As briefly indicated above, Messick's chapter on Validity is a chamber of horrors, a gruelling journey through deep and varied sources of invalidity that would surely deter any rational person from ever attempting to show that any test was valid. Yet again and again he slides back into psychometrics, into "multiple choice" tests, into technological fixes, into the fudged solution.

Here is one such: "Tests," explains Messick, "are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both"(p34).

Not so. Messick has, conveniently, left out the fourth alternative, "or neither." And surely this is the alternative most congruent with his own analysis. By doing this he has assumed the very thing that is in doubt - that the construct can, in fact, be measured at all, in the light of epistemological issues, multi-dimensionality problems, value confusions, comparability errors, and so on.

Summary

To summarise, the notion of error is circumscribed by the construction of the event being described, just as it is boundaried by the epistemological assumptions of the judgment process.

Theoretically, error in assessment contains within its ambit all those ontological inadequacies, all those epistemological slides, all those logical contradictions, all those semantic obfuscations, all those definitional fudges, all those ideological camouflages, all those value variations, as well as all those potential empirical falsifications of implicit truth and accuracy claims, that characterise the field.

Practically, the description (measurement) of error is not dependent on any notion of a single truth, but rather on one of differences between multiple truths, all with some claim to legitimacy; these are implicit in the production of the assessment event, in the interpretations of the assessed and the assessor's experience of that event, including categorisations, and in the particular intended and received meaning of the communication of that judgment to others. The error becomes explicit when all of these phases of the assessment event are pluralised; when genuinely independent events are constructed; when independent categorisations are produced by participants in the event; when the judgments, and the meanings given to those judgments by involved persons, are compared.

Conclusion

Thus whilst the theoretical aspects of validity may indeed be fully discursive as Cherryholmes (1988) argues, the practical extent of invalidity is demonstrable as an empirical reality in the material world, partly as a result of that very discursiveness. For example the analysis presented earlier of the electrical automotive test presented irresolvable complexities in determining what empirical meaning could be given to the validity of the assessment. As the notion of validity is currently constructed, it would be resolved, if it was attended to at all, by the validity advocate giving an expert and coherent case for the defence, which would be unchallenged. That is, it would be resolved by resort to the Judge's frame of reference, and ignoring the other frames.

From the standpoint of invalidity, there is no such confusion. All of the suggested measures are useful measures, and the range of estimates that they produce for any one trainee indicates the range of error within which that person is being categorised. And we should not be surprised if at times this range covers the whole range of categories available.

As indicated in earlier chapters, the categorisation of persons has enormous effects on people, both in terms of their conceptions of themselves, and in their subsequent implicit and explicit exclusion from occupational opportunities. Such exclusion is not a discursive practice, but a very practical reality, though doubtless language is a significant factor in the acceptance of the violation. Further, the immense uncertainties associated with such categorisations is both demonstrable and measurable.

I have argued that validity discourse is currently constructed in such a way as to deny this demonstration. Invalidity discourse, based on the detailing of error components as presented here, is an advocacy for the defence of the examined rather than the examiner. As such it tends to redress the power imbalance, and hence reduce structural violence and increase social justice.

Part 6: Application

Chapter 18: Competencies, the great pretender

Synopsis

In this Chapter, I apply the philosophical and conceptual positioning, tools of analysis, and the reconceptualised sources of error developed in this thesis to the competency based assessment policies and practices of Australia in the 1990s.

I first indicate how the notion of competency standards is overtly central to the whole competency movement, the introduction of which is shown to be overtly politically motivated. Thus the crucial links between political power and educational standards that are argued for in Chapters 3 and 4 become transparent.

I then go on to examine the validity, or more accurately, the invalidity of competency standards in the light of the thirteen sources of error specified in the previous chapter. The applicability of this analysis to a particular case is thus demonstrated.

Context: The re-birth of competencies in Australia

In the 1980s the discourse of politics became subsumed within the discourse of economics; quality of life was implicitly submerged, becoming a by-product of standard of living. And standard of living was explicitly defined by empirically derived statistics selected and interpreted by the theory of economic rationalism. Thus were the concomitant subjectivities of human misery, and the appalling atrocities of environmental degradation, excluded from the mainstream debate.

This same movement saw management practices move from control and exploitation of workers, modified at times by paternalistic concern for them, to a set of more manipulative practices described under the rubric of human resource management. This required the objectification of workers as a set of competencies, a necessary precursor to their ultimate replacement by more efficient computerised and robotic systems.

So the imposition of competencies as the basis of Australian technical and professional training during the late 1980s was in no way a decision informed by considered professional opinion; it was, from the start, an overtly political manoeuvre designed to solidify economic ideology in work practices, to demonstrate how skill and efficiency would reap their rewards in the "fair and just" game of the new internationally-competitive capitalist world order:

> The National Training Reform agenda is a co-operative national response to economic and industry restructuring, including labour marked imperatives and emerging requirements arising from workplace reform. The overriding aim is to increase the competitiveness and productivity of Australian industry, through industry responsive reform of the vocational education and training system. Flexibility to meet enterprise requirements within a stable and consistent national system is essential (National Training Board, 1992, p4).

The report goes on to state that "National competency standards provide the focal point of the new competency-based system" (p4). So here, quite explicit at the heart of the system, the manifest pivot, is the ubiquitous standard.

And of the essential arbitrariness of those standards or the necessary error in their measurement there is no word in this seventy one page report. Those two pillars of educational measurement, reliability and validity, do get a mention in the last page of the report. We are informed that assessment under the National Framework for the Recognition of Training (appropriately capitalised as a recognition of omnipotence) "provides for consistency as well as quality," and that one of the five principles of this approach is that "Assessment practices used shall be valid, ie. the techniques used must actually assess what they claim to assess." Furthermore, they must be reliable, in that "assessment approaches shall be able to be relied upon" (p71). And the Lord said, "Let it be done." And behold, it was done.

As Mc Donald (1994) nicely puts it: "The piece of commonsense that says that merely categorising something does not necessarily mean you can measure it easily, seems to have been lost" (p2).

In the first six pages of the NTB report, dealing with overview and context, the word "flexibility" appears eight times and "consistent" or "stable" six times. It is within this fundamental tension that the whole framework contradicts itself into nonsense. Let's unpack the argument in some detail:

> There is a clear need for a stable framework for national competency standards which is consistent across industries and across Australia. . . enabling nationally consistent assessment and certification to be achieved over time (p8).

Who needs this is unclear, but the rest is clear enough. The framework, which includes the ontological and epistemological assumptions about skills and learning and knowledge and the axiological assumptions about value, as well as the frame of reference about assessment, are all to be imposed on the basis of some "need." In other words, a centrally controlled system of education and training is to be imposed.

But there is to be flexibility. "Flexibility is required to enable specific industry and enterprise characteristics and necessary performance outcomes to be accommodated" (p8). The report goes on to indicate what is meant by flexibility; how flexibility is itself to be stabilised:

> This flexibility will be facilitated by the inclusion of general skills and knowledge in industry standards. Ensuring that industry standards look to the future, are packaged to allow multi-skilling, concentrate on important common skills and, where possible, not tied to particular forms of work organisation . . . simplicity as well as flexibility (p9).

In other words, flexibility is to be achieved by making the competence standards both general (non-specific) and simple.

So what have we got here? National Competency standards "provide the specification of the knowledge and skill and the application of that knowledge and skill to the standard of performance required in employment" (p9). And "assessment is the process of judging competency of an individual against prescribed standards of performance" (p11). So assessment is clearly in the Specific frame of reference, related to specific workplaces, related to particular jobs, related to performances that can be specifically described, and their levels clearly delineated and categorised (Norris, 1991).

On the other hand, the competencies are to be general, are to be non-specific, and furthermore have the truly remarkable quality of reflecting "not only industry's current but future needs"(p8), indicating perhaps a growth industry in astrology and clairvoyance training.

That the contradictions are so explicit gives hint to their genesis; they are the product of a succession of committees: Special Ministerial Conference on Training (1989), the Finn Report (1991), the Carmichael Report (1992), the Mayer Committee's Report (1992), and finally the input of the committees of the National Training Board.

At what level of discourse is all this? Are we involved in discourse at the rational-empirical level? Is this rhetoric really about measurement of competence? Or is this discourse at the mythical level? Is this about the construction of a national icon called competency standards around which a whole structure of power relations may be developed, and a whole new generation of workers constructed?

Porter (1992), speaking more specifically of the Carmichael report, sees it as

> a clever piece of policy writing, since its emphasis on diversity, options, pathways, and so on, obscures its desire to develop a training structure that is uniform, standardised and under the control of centralised bureaucracy (p54).

Similarly, Jackson (1992) is concerned that

> all of these reforms can be seen as a process of ideological capture, replacing the public purposes and social vision of education and other social institutions with the logic, and the social relations of, private wealth creation. The result is a profound and fundamental shift in where and how, and in whose interests, these institutions are controlled and managed(p159).

This is not to uncover a conspiracy to disenfranchise learners, teachers and small employers; Beevers (1993) explains that:

> In fact the Labor Party and the union movement in particular appear to have set out to do exactly the opposite. However the adoption of positivist, rationalist, bureaucratic and corporate managerial values and procedures has given rise to a curriculum model that - while providing advantages for politicians and systems managers - discriminates against the learning process and hence teachers, learners and small employers. . . What has been silenced is knowledge and skills that do not fit the technocratic, scientific, rationalist paradigm. The only knowledge and skills deemed worthwhile possessing are those believed to be directly related to increasing economic productivity (p103).

The paths to violence are indeed paved with good intentions.

Sources of error in competency assessment

In the remainder of this chapter I shall examine competency assessment in terms of the thirteen sources of invalidity conceptualised in Chapter 17.

Temporal errors

Firstly, there are the temporal errors in the criteria themselves. As Melton (1994) comments, "Inevitably the standards set reflect the perceptions of a particular group responding to perceived needs at a particular point in time. These will change as perceptions and needs change with time"(p288). Perhaps errors of this type are best categorised under construction errors, which become solidified in time because of the enormous structural and bureaucratic complexities involved in the production of the criteria of competency, which, despite what Melton says, do not usually specify a standard, a measurable level of adequacy.

Of more immediate concern are the variations in a particular person's performance over time. How are these to be interpreted? If competency is to be attached to the assessed person, then one adequate performance means the person has it, so is competent, so long as we assume that the thorny problem of adequacy has been solved. But if no adequate performance occurs, this could be a function of context, and does not necessarily imply incompetence. On the other hand, if competency is attached to events, then every performance ought to be adequate if the person is competent. Regardless of context? Confusion abounds!

In practice, temporal errors are confounded because no two events in which a human can engage can be identical, because time is change. Just as no two evaluations of a human event can be identical; they may involve identical categorisations, but the interpretative meaning of those categorisations change with time and with persons.

Potential confusions around the temporality of the measurement of competency standards abound; in current practice they are solved by pretending that they do not exist. They are a major source of error and confusion related to the meaning of any such measure.

Contextual errors

If performance depends on context, then assessment is about events in context, and competence is someone's judgment that a particular contextual behaviour is adequate. This is surely what "work" is all about, whether it is in school or on a personal project or in a paid job.

But this is so messy, because then a label can't be pinned on a person, because it belongs equally to a context. So how do we get back to a context-free categorisation? Easy:

> the assessment of competence is fundamentally about inferring competence from samples of performance. Under these circumstances, "competencies" are defined in terms of attributes, the competence is seen as deriving from the possession of, and ability to apply, relevant attributes to occupational tasks (Bowden, 1993, p55).

Of course, if fundamentally the assessment of competence is about inferring from samples of performance, then that's what you do, and the stuff about attributes is irrelevant. The attributes are politically necessary to get rid of the contextual error by assuming that there is none, so that the person can be categorised for all contexts.

Unfortunately, the empirical data contradicts the assumption, and makes the idea of such "attributes" very suspect, or at the least quite unmeasurable. Stanley (1993) sums up the current position:

> The message from the literature on transfer of training is that the idea of general strategies or competencies has been oversold. There are no substitutes for the building up of knowledge bases in specific domains. The evidence emerging from a number of recent cognitive studies is even stronger. It suggests that ways of thinking applicable for one domain of knowledge may be inapplicable in another (p145).

So the cost of attaching the categorisation to the person is to make it invalid in real contextual situations. The notion of competency standards solves the issue of context by fantasising the notion of an "attribute" called "competencies" which belong to the person so are independent of context. Reliability is thus increased in a psychometric sense. And validity is greatly diminished.

Construction errors

The original idea of competencies, in the Specific frame of reference, was to detail and teach all the little tasks that seemed to constitute the performance, and then test that they were all learnt to the required level of adequacy. The notion of competency standards as currently interpreted has moved a long way from this reductionist view. As Bowden (1993) explains it:

> the approach being taken to develop competency standards for the professions in Australia is not based on the professional's ability to perform specific tasks, but on the integration of relevant knowledge, skills and attitudes to complex workplace activities (p54).

Based, that is, on the measurement of knowledge of doubtful applicability and relevance, of skills that certainly have different applicability to different contexts, and of attitudes about which any inferences are surely problematic, and any measurement is highly suspect. So the price of solving the reductionist tiger has been to create an overgeneralised, undefined, unmeasurable and mis-attached elephant.

Melton (1991) elucidates the dilemma cogently:

> If competence is thought of as a deep structure of general ability then it is difficult to see how this abstract construct can be related to practice. It is close to offering a general theory of intelligence in forms of cognitive potential (p334).

It does indeed, and such a route is very rocky, as the last hundred years of controversy about intelligence tests have indicated.

Labelling errors

There are, as in all assessment systems, two types of labelling errors: There is the label of the particular competency; and there is the label of the categorisation of that competency.

As we move away from the Specific frame that can describe very specific eventful behaviours, we experience greater confusion in the meaning of the name that will become, in discourse, the referent for some practical competency that is ultimately defined either by some practical events in the world of work, or as some attribute or trait of a particular person. Regardless, whether we are talking of very generalised competencies such as "understands basic scientific principles," or very specific ones such as "adds two 2-digit numbers," what competency might mean in these domains is inevitably contested, and is different when viewed from different value positions or contexts, so the name will mean different things to different people. And this is not solved by the curriculum or test writer redefining such terms for their own purposes. As explained elsewhere, such a tactic may increase the reliability of the test, but it also increases its invalidity, because the user of the data generated from such tests is necessarily constrained to interpret the data in terms of the labels provided; labels to which they will attach their own meanings for their own purposes, and not magically absorb those constructed by remote curriculum and test constructors.

Similarly for the meaning of the label "adequacy" which is a necessary component of any discourse about competencies. Even if the problem of the meaning of the label that describes "what is being measured" could be solved, and we had a "scale" that was valid, we are still left with the problem of what is adequate along that scale; with the problem of the standard. This is also permeated with arbitrary and idiosyncratic definitions and interpretations, as well as enormous contextual variations; in short, another immense area of uncertainty, confusion, and hence error in personal categorisation and its interpretation.

Attachment errors

When competencies are described in terms of some particular assessor's evaluation of "adequate" work performance in a specific workplace, attachment errors are at a minimum - so long as competence is clearly tied to that particular work at that particular place by that particular person. Any reduction of the specification description, any attempt to attach the label to the person assessed, represents an attachment error, and, at least in the philosophical frame of this study, makes any competency claim ontologically invalid.

When such competence is reduced to a number of specific performances under specified conditions at specified levels of adequacy, attachment errors are at a minimum when all of this information is retained in the assessment description. Attempts to combine this information into one statement about competency, of which the specific behaviours are elements, is a logical type error which makes any competency claim logically invalid. Attempts to give a meaning to such a summation of elements involves both a comparability error, and an epistemological error in that the summation can have no meaning. Any such summation, by losing the contextual data related to the individual elements, results in an attachment error because the data now becomes attached to the person being assessed.

When, on the other hand, competence pretends to be some fixed attribute or skill or trait of the person examined, an attribute that is somehow "measured" by the person's interaction with a test, then the attachment error occurs when this measure is attached to other contexts, to other workplaces. It will then become apparent as contextual error or prediction error.

Frames of reference errors

Already the instability of the concepts of "competency," "competencies," and "competent" have been demonstrated. Norris (1991) comments that

> The requirement that competencies should be easy to understand, permit direct observation, be expressed as outcomes and be transferable from setting to setting suggests that they are straightforward, flexible and meet national as apposed to local standards . . . as tacit understandings of the words have been overtaken by the need to define precisely and operationalise concepts, the practical has become shrouded in theoretical confusion and the apparently simple has become profoundly complicated (p331).

He goes on to explicate:

> Behavioural constructs . . . express what is to be learnt in ways that make it transparent, observable and measurable. In contrast . . .the generic competency approach defines competence as broad clusters of abilities that are conceptually linked (p332).

In other words, the behavioural construct of competence is in the Specific frame, and the generic is in the General frame.

Messick (1984) confuses the issue further when he claims that competence is what a person knows and can do under ideal circumstances, whereas performance is what is actually done under existing circumstances. So competence is potential, is ability imminent. It follows that one successful performance demonstrates competence, because the conditions cannot be more than ideal, so one must assume they were less for any successful performance. On the other hand an unsuccessful performance can never demonstrate incompetence, because the conditions may not have been ideal.

So in theory there is confusion as to whether we are dealing with traits or demonstrated skills, concepts that require the General frame of reference, or particular defined behaviours, which require the Specific frame. In practice the confusion proliferates, for invariably the description of the standards that define the cut-off is either non-existent or vague, as indeed is the measuring instrument or instruments which will provide data to which the standard must be compared. So the practical assessment of what is adequate must be made in the Responsive frame - an intuitive response from the assessor. Such "subjective" admissions are, of course, utterly inadmissible, for the success of the whole charade is dependent on the appearance of objective accuracy and precision. Luckily, this is possible if the assessment mode shifts to the Judge's frame. So this is what happens, and certainty is reestablished, albeit in a different frame than that theoretically intended.

To summarise, analysis of competency assessment in terms of frames of reference indicates semantic chaos, discourse riddled with self contradictions. Out of it all there still emerges, from all involved, belief that the system works.

And in as much as people are categorised, it does indeed work. To further believe however that some accurate measure of minute error has emerged from such conceptual confusion and personal lack of awareness is to substitute blind faith for rational thought.

Invalidity from this source is thus profound, and stems from the epistemological irrationality that must occur when frames of reference with contradictory assumptions are amalgamated without distinction into a single discourse.

Where does all this leave the individual student? Apparently presented with a list of clearly defined outcomes, things to know and do at predetermined levels of competence, closer inspection leaves the student with a list of ambiguous topic headings and ill-defined "skills," on the basis of which he or she will be tested, and then categorised by comparison with opaque standards visible only to the professional eye of the teacher. Was it ever otherwise?

Instrumental errors

Referring to standardised and/or criterion referenced tests, Berlak (1992) notes that "The credibility of these tests depends upon the claim that they are scientific instruments" (p181). Just so the credibility of competency assessment as a whole. The notion that these assessment systems are based on the measurement of clearly defined standards is what provides the educational, moral and public relations glue that transforms a set of fragile value and assumption struts into a powerful cognitive structure.

Yet it is surely a false claim. There are rarely such standards available, even at the practical level. At the level of physical factory products, standards that are related to some criteria of quality can sometimes be set up and measured, but these are a far cry from the "attributes" that predate competence in personal performance.

As described in the section on frame of reference errors, the whole discourse is emersed in epistemological confusion. What is important to note here, however, is that by pretending to belong to the Specific frame, the professional necessity to provide estimates of standard errors of measurement, necessary in the General frame, is side-stepped; not that educational practice ever paid much attention to that professional necessity.

The instrument, as apposed to any test, thus is firmly inside the mind of the assessor, an intuitive judgment hidden beneath the overt scientism of the competency label with its overtones of specific behaviours and definable standards.

Categorisation errors

Competencies must be described and then categorised. To categorise a competency we must first measure it and then compare the measure to a standard. As a result of this comparison we may then categorise the performance as adequate or inadequate, or the person as having, or not having, the competence.

So can we measure accurately these competencies that are described? Norris (1991) comments :

> there is a massive mismatch between the appealing language of precision that surrounds competency of performance-based programmes and the imprecise, approximate and often arbitrary character of testing when applied to human capabilities (p337).

As to the standards, these are normally presented as criteria to consider, as hints to decision makers, rather than defining the point on the measure that dichotomises a continuity. And even if there was a scale or measurement, and so the "standard" could be specified, how could it ever be anything other than arbitrary? A political decision based on data permeated with individual subjectivity and value.

Levin (1978) described the use of minimal outcomes in schools in the United States. It applies equally to the use of competencies in Australia in the 1990s:

> we do not have the knowledge bases to construct a defensible set of performance standards for certifying student competencies except in the most arbitrary sense. Whether such arbitrary standards are worthwhile in themselves may be debatable. Their inability to predict with any confidence that which is important in adult life is not debatable (p314).

Comparability errors

Melton (1994) accurately describes the sort of processes that are actually involved in competency assessment:

> Assessment is not simply a matter of ticking off whether individuals can or cannot perform tasks to certain clearly defined levels. Rather it is about looking at evidence, and making judgments about the levels of competence achieved based on the evidence provided. The evidence may be gathered from a variety of sources including observation of performance in the place of work, observation of specially set tasks, records of tasks that the candidate has performed in the past and from questioning the candidate on any aspect of the performance. Clearly much judgment needs to be brought to bear in interpreting such a range of evidence (p288).

And, of course, a judge will give a particular weight to a particular source of evidence, and will give a particular interpretation to the data available from each source, so that the meaning of any such final judgment must be quite obtuse, and different from the meaning given by another judge, even if the categorisation is the same, which seems unlikely in most cases.

Prediction errors

Because competencies in Australia have been specifically politically invoked to improve work practices and hence profitability in industry, prediction errors occur when the produced competencies do not specifically do all of those things; it is possible, remotely so in my view, that the educational events wrapped around competency standards might indeed in some cases have some validity in regard to the first of these claims, related to work practices, though some early research does not support this (Gillis, 1995). Of course, even if there is some correlation between the competence measure and some later predicted outcome, this in itself does not indicate causal link between the two categorisations that is mediated through the competency attribute.

In fact, as I have argued in the section on Consequential errors, it is unlikely that any empirical data will be collected in this regard because it is the assumption on which the whole scheme is premised, and thus not amenable to investigation.

Logical type errors

In all of its cyclic incarnations, competencies as specific behaviours have invariably encountered the criticism that they are reductionist, that they fragment knowledge, that they are in essence, trivial. Perhaps it is sufficient here to give two references:

> It cannot be assumed that mastery of the elements of competence will automatically lead to the achievement of more complex skills in the higher reaches of the hierarchy (Melton, 1994, p188).
>
> If I were to place competence within the art of pottery which I practise. Seeing it wholistically from the perspective of a great tradition of planetary and historical scope, I would only say: competence, your name is mud. (Beittel, 1984, p119).

In the Australian context, competencies face an identity crisis in that they are uncertain whether they are to be interpreted as holistic summations of such specific behavioural elements, or as specific behavioural outcomes of holistic mental attributes.

If the former, then the logical type error occurs in the summation, in the confusion of members of a class (the specific behaviours) with the whole class (the competency). In the latter case the logical type error occurs in the confusion of a description of a class (the generic competency) with members of that class (specific context-related work performances).

Either way confusion is confounded and error escalated through the attempt to define and describe competencies in any place other than their area of actual performance.

Value errors

Competence implies some purposeful act; a person is competent when she does something adequately in some context. So the first question to be asked in a competency judgment is: What ought the person do in order to be deemed adequate? This is not a factual question, but a value premise. And it is where every list of competencies must begin. Pearson (1984) argues that "until the value premise is made the competency claim cannot get off the ground" (p34). Thus all competency descriptions are based on value premises, which are usually unstated.

One implication of this is, as Norris (1991) points out, that "In the effort to describe competence in precise, transparent and observable terms, to predict the specific outcome of effective action, what is in fact happening is the pre-determining of good practice" (p334).

To the extent that competency requirements dictate school programmes, they also determine that "The measure of success that is applied for the schools is not the degree to which they foster intrinsically meaningful activities, but the degree to which they satisfy competence-related outcomes (Levin, 1978, p311). Levin (1978) goes on to assert that "Certification standards are signals to the schools of what is considered important by society, and their message will not be lost in individual teacher decisions or organizational ones" (p314).

Jackson (1993) perceives that the underlying intent of competency based teaching and assessment is to provide more governmental control on teaching institutions, and any effect on individual learning is secondary to this:

> the achievement of competency-based curriculum may not be about lasting improvement in individual performance at all, but about making teaching and testing accountable to a standard through a warrantable set of procedures. Technically, it is not the competence of the individual which is assured by these methods, but the competence of instruction and the liability of the institution. The shift is central to the power and sophistication of the competency paradigm as a tool of governance and an ideological force (p157).

How are these values transmitted to the individual student? What is the value learning that accrues? Here is a world of learning presented with machine-like crispness, sets of facts and relations and skills as neat as a computer board; the world of learning and of work reduced to packaged modules to be eaten up and deposited in the appropriate mental filing cabinet for later reproduction at so many dollars an hour.

Yet as we have seen, this whole operation begins from a particular view, generally not stated explicitly, of best practice: a particular positioning; a particular attachment to certain sorts of power and affect relations; a particular consciousness about work and its effects; and a begging of the question of who benefits from these particulars.

Where does the individual student position himself in this value matrix? He is supposedly acquiring the competencies that will allow flexibility in various job performances. Yet his experience may deny the usefulness and relevance of what is being presented. Even so, the competencies must be achieved. So rather than flexibility, such a student will learn not flexibility, but conformity; not a producer of new work practices, but a consumer of old ones.

Invalidity in terms of value then derives not only from the bias that derives from unstated value assumptions, but from the very specificity of stated intentions, and their contradiction by associated social effects; that is, by those very contradictions that are at the heart of symbolic violence.

Consequential errors

Elsewhere in this dissertation I have argued the centrality of assessment procedures to the construction of the individual in society. Commenting on the scene in the United States, Berlak (1992) comments:

> Among all assessment procedures, standardised and criterion-referenced tests are particularly privileged, that is, they serve as the single most powerful regulators of schooling practice, shaping the language used in public discussions about schooling, the criteria for judging the competence of students, and the range of possibilities considered for reforming the schools (p194).

And Jackson (1993) sees Australia in the 1990s following along a similar path:

> the discourse of competency increasingly defines not only our current practice but also the parameters of our imagination on issues of education and training (p159).

So here is one clear consequence of the competency movement. Increasingly the boundaries of discourse become narrower, and the possibilities for diversity become constrained, as notions of specifiable behaviours, performances, outcomes, skills and abilities, all defined by persons outside the training institutions, begin to dominate educational discourse. There is the further mythical belief that in some magical way standards are incorporated into these competency descriptions, which can be precisely measured and compared to such standards.

Students in this context are cogs in a gigantic machine. They are disempowered in terms of the substance and the value assumptions that predate what is to be learnt. There is no notion here of learning that grows out of specific purposes, learning styles or values of students, or of curricula negotiated to meet such purposes. Nor indeed is there any sense of relatively autonomous teaching agencies offering, among them, a proliferation of solutions to the relatively intractable problems of job training. As presented, competency assessment is the solution. The problems, whatever they may be, have been pre-empted. The job of training is to implement the solution. The function of evaluation is to indicate that the solution has been implemented. The closed black and white fantasy circle is complete.

Invalidity in terms of consequences stems most profoundly from the loss of the initial problem, which has been firmly removed outside the closed circle of competency discourse. For the National Training Board (1992), "the overriding aim is to increase the competitiveness and productivity of Australian industry"(p4), an aim now subsumed under the solution, which is assumed to be the National competency assessment system; so within the discourse of competency assessment not only can this question about productivity not be answered, it cannot even be asked, because its answer is itself.

For the individual student the potential errors in categorisation are immense. This particularly applies to students already enmeshed in work practises. Their learning cannot be based on local analyses of work environment deficiencies, or on creative transformations of work practices, because it is dedicated to their absorption of pre-ordained competencies which are supposed to magically provide such solutions. And if (when) the magic doesn't work, there is no place to go, for success has too long been dependent on the acceptance of absurdity.

Summary

I have argued that there are at least thirteen sources of invalidity that affect the measurement of competency standards. I contend that any one of these, applied to the assessment of individual students, would make the assessment of that student in these terms invalid.

Chapter 19: National tests and university grades

Synopsis

In this chapter I apply the reconceptualised notion of invalidity to national literacy testing, and to the definitions of grades within my own university.

These are presented as specific examples of the potency of the invalidity conceptualisation.

National Literacy Testing

Context

In its edition of 15-16 March, 1997, the newspaper Weekend Australian announced on the front page under the heading "All pupils face tests of literacy" that:

> The literacy and numeracy of every Year 3 and 5 student will be tested from next year under a historic agreement between the Commonwealth, States and Territories yesterday.
>
> The Catholic and independent schools sectors have indicated they will support the national testing program, which will be linked to uniform education standards to measure the reading, writing, spelling and mathematical ability of students.
>
> . . . The federal Minister for Schools, Vocational Education and Training, Dr Kemp, described the literacy strategy as a "historic agreement for the children of Australia" because it stresses that every child starting school from next year will be able to read, spell, and add up within four years.

The literacy test is to be based on that developed some years ago by the NSW Education Department, and it is this test to which the following critique is addressed, in terms of the thirteen sources of invalidity.

Temporal errors

Temporal errors are indicated by the differences in assessment description when the assessment occurs at different times.

No estimates of temporal errors in the national literacy testing program exist. They would, of course be easy to obtain and would be small compared to some of the other sources mentioned here. Small, that is, for most students. But the same theory that predicts this also predicts that a small percentage of students (randomly placed and unfindable) would have large discrepancies. But even small discrepancies would destroy the notion of infallibility that seems to be necessary for such tests to be publicly acceptable. This is what test administrators call public confidence, and I have more accurately named a psychometric fudge.

Contextual errors

Contextual errors include all those differences in performance and its assessment that occur when the context of the assessment event changes.

Literacy is a concept of great educational importance, of diffuse and contested and multi-dimensional meaning. It involves at the very least reading and writing. Yet reading what under what conditions? And writing what under what conditions? A test defines the what and defines the conditions: Tightly specifying the conditions improves the reliability; yet at the same time it obviously disguises and increases the lack of generality and hence increases the contextual invalidity.

Essentially, the context of test-taking is not the context in which literacy, in most of its forms, is demonstrated.

Construction errors

Construction errors are indicated by all those differences in assessment description when the same construct is assessed independently by different people in different ways, whilst the broader context of the assessment is held constant.

It would be relatively simple to take samples of children and have teachers and researchers and the children themselves make independent assessments of various aspects of their literacy, and estimate construction errors by comparing the estimates with each other and with the result of the test. This writer has no doubt that such an experiment would presage the immediate cessation of such testing.

Labelling errors

An assessment must be an indicator of something. It must have a name. Differences in the meaning of the name, both before and after the event, constitute confusion and hence error. Labelling errors are defined by all the differences given to the meaning of the assessment (what it actually measures) by all the participants in the assessment event(s), and by the users of the assessment information.

Literacy tests presume to measure literacy. But which particular aspects? What could any test score tell us about any of those aspects? What meaning is given to those aspects by any particular teacher? How does that meaning compare to that teacher's concept of literacy? And what action could be taken by any such teacher on the basis of those meanings to help any child more than that teacher is currently helping? The extent to which these questions produce diffuse and varied and contradictory answers gives an indication of labelling error. And the meaning of literacy includes such confusion. The problem is not solved by imposing a definition; this enables us to increase reliability, and reduce the apparent error in measurement. But it is a reductionist trick, a semantic scam. The concept of literacy is diffuse, so any attempt to measure to is, at best, extremely imprecise, and, at worst, meaningless and hence impossible.

Attachment errors

Attachment errors are the ontological slides that occur when a description of a relational event is attached to one of the elements of that event; specifically, when a complex relational event involving the construction of a test, an interaction of the test with a person, and a judgment of an assessor, is described as a property of the assessed person, this is an error in attachment.

The implications of this source of invalidity for literacy testing are immense. Any information about the test cannot be unattached from the particular test and attached to the student as a "trait" or "ability." This involves a demystification of the whole process and its highly suspect theoretical underpinning. Such demystification relates it to the fundamental question "What do we really know about where this literacy score came from?" The answer is clear. A particular group of people selected a particular set of multiple choice test items which the student answered under particular conditions and were subsequently given a score which placed them in a rank order and some of them were then classified as below a standard which did not exist until this group or another group were so classified.

The point to emphasise here is that the score does not belong to the student. It belongs to the experimental event of which the student was a part. Any movement beyond this point requires another experiment - which, of course, produces another event, with concomitant multiplication of confusion and error.

Frame of reference errors

Practically, frame of reference errors are indicated by specifying the frame in which the assessment is supposedly based, indicating the errors according to its own and other frames, and indicating any slides or confusions that occur during the assessment events.

In testing programs on literacy the tests pretend to be in the Specific frame of reference. The tests are talked about as though there are clearly defined and accepted specific tasks which students must do successfully in order to be considered literate or numerate. And that there is some predefined standard to which appeal may be made. Neither of these claims are true. The test items which are the basis of complex statistical manipulations are subjectively chosen by test constructors from the pool available, which may include some that they themselves specifically construct. And there is no standard other than that defined by the test itself. Some test constructors talk of an absolute scale. They are deluding themselves (Behar, 1983). All test data are based on item statistics which are norm referenced from groups of test takers. So the tests produce a rank order of merit and the test controllers (test makers, educational administrators, or political funders), make arbitrary decisions about adequacy (Glass, 1978). What we can be certain of is that the tests will produce a rank order in which some students will obtain higher scores than others. That is what the tests are designed to produce. Any implications beyond this about adequacy are arbitrary value judgments.

Instrument errors

Instrumental error is implicit in the construction of the measuring instrument itself; what is conventionally called standard error of the estimate, or is indicated by the spread of judgments of independent assessors about a particular performance on a particular test.

One assumes that in national literacy tests this relatively small source of error (simple reliability) will be known to test constructors, forgotten by test administrators, and withheld from teachers and students. Regardless, such an estimate of error gives no information about the error of a particular student, and withhold the statistical information that only two thirds of actual students will have "true" scores within these limits, and as the total numbers tested increase, an increased number of individual students will be given completely unacceptable estimates.

At a more fundamental level, the instrument (the test) cannot measure anything because there is no Standard, no adequate theory -practice bridging to define the scale, no scale, and thus no measure that the scale may proscribe, that may subsequently be compared to a standard of acceptability.

Categorisation errors

Categorisation errors derive from confusions about the definition of standard of acceptability, from differences in the meaning of what is being assessed and in the magnitude of its measurement, and in the variability of the judgment process in which the comparison with the standard is made.

Practically, categorisation errors are all those differences in assessment description that occur when particular data is compared with a particular standard to produce a categorisation of the assessed person.

The implications of this for literacy testing are profound. For not only is the meaning of the score highly suspect, but there is in fact no standard of literacy with which such a score may be compared. The standard is an arbitrary point selected after the event by the test makers and is based on the particular test, or on the particular items used in the construction of the test. Such circularity in definition produces a closed system that is the stuff of fantasy, but not of scientific measurement.

Comparability errors

Comparability errors include all those confusions about meaning and privileging that inhabit the addition of test items, test scores or grades. Practically, comparability errors are indicated by constructing different summaries or summations according to competing models. The differences that these produce indicate the comparability error.

Literacy is a multi-dimensional concept. As such, a single dimensional scale can be used to measure the concept, but such a measure could not be given a meaning. In particular, any categorisation (involving a standard, assuming one exists) cannot be given a meaning, because it could never be certified whether any particular single - dimensional score was above or below that "standard." Because such meaning is central to the notion of validity, such inability to give a meaning makes any uni- dimensional test of literacy constitutionally invalid.

Prediction errors

Practically, prediction error is indicated by the differences between what is predicted (or more subtly implied) by the assessment data, and what is later assessed as the case in the predicted event.

There is an implication in the national literacy program that the scores show that some children are illiterate, and that without special intervention triggered by this test they will remain illiterate. Such an implication could be empirically tested, assuming there was some satisfactory definition of illiterate. I know of no such definition, or of any program to develop one or otherwise empirically test the effects of the testing.

Logical type errors

Logical type errors occur whenever there is confusion between statements about a class of events, and statements about individual items of that class. Practically, logical type errors are made explicit when the explicit and implicit truth claims of a particular assessment are examined and any logical type errors are made explicit. Such exposure may invalidate such claims.

In a rare burst of intellectual honesty the earlier versions of the literacy test were headed "Aspects of literacy" (NSW, 1995). Such a test cannot be a test of literacy. Statements about some members of the class do not apply to the whole class. All literacy and numeracy tests have this problem. They are essentially a summation of the specific items that the test comprises, and assumptions cannot be made of implications beyond this. Psychometrics could be defined as a statistical sampling game that produces a fantasy about traits in order to sidestep the contradictions that flow from the reality that all test scores are summations of discrete elements, and that all information about the individual elements is lost in the summation.

Value errors

Value errors are indicated by making explicit the value positions explicit or implicit in the various phases of the assessment event, including its consequences, and specifying any contradiction or confusion (difference) that is evident.

The National tests purport to give information about individual students that might lead to remedial action. The value appealed to is that of helping students and improving performance. The tests are not diagnostic and so give no information about what particular misconceptions or problems (if any) particular students may have, apart from the extremely error-prone response from one or two items. Even if such diagnosis were available, its usefulness would depend on teachers being able to use it to improve student performance. And since it is not known whether or not teachers have already targeted some children for extra attention, its usefulness would depend on whether the test produces the same group for special attention, and in cases of difference whether the National test produced a more valid selection.

As there is no evidence that the tests will help children, it may be less naive to suggest that the main value behind the test is to help politicians gain prestige by appearing to solve a problem (which may not exist).

Consequential errors

Consequential errors are indicated by the differential positive and negative effects that individual teachers and students attribute to the assessment process. At a more profound level the test may involve an explication of the very construction of their individuality, and all of the potentially violating consequences of such constructions.

The focus of the testing will be on those who are lower in the rank order. Theoretically these will be identified, and will improve as a result of special instruction. The magical improvement kit has not yet been produced, so such consequences are doubtful, especially as literacy (as most people understand the term), is so dependent on a whole range of experiences outside the school. What is more certain as a consequence is that such students will be classified as "failures" or "remedial" and will, in many cases, construct their individuality accordingly.

Summary

Practically, the description (measurement) of a person's literacy is not dependent on any notion of a single truth, but rather on one of differences between multiple truths, all with some claim to legitimacy; these are implicit in the production of the assessment event, in the interpretations of the assessed and the assessor's experience of that event, including categorisations, and in the particular intended and received meaning of the communication of that judgment to others. The error becomes explicit when all of these phases of the assessment event are specified; when genuinely independent events are constructed; when independent categorisations are produced by participants in the event; when the judgments, and the meanings given to those judgments by involved persons, are compared.

When such errors, contradictions, and confusions are acknowledged, the pristine purity of the test score disappears, to be replaced by a wide fuzzy band of possibilities; then rank orders recede, standards evaporate, categorisations are exposed as fantasy, and the whole inane and monstrous structure crumbles.

National literacy tests have thirteen charges (at least) to answer before being considered valid. Many of these are so fundamental that I doubt any reputable educator would take the case.

University grades

Context

Just as honesty begins with self, so truthfulness should not ignore the home campus. My own university has announced a new grading system for the categorisation of students (Flinders University of South Australia, 1997). An analysis of the grade descriptions indicates six criteria are used. A summary of the descriptors is given in Table 1 (see next page).

In the next section I will examine this grading system in terms of the thirteen sources of invalidity.

Temporal errors

If grades refer to a particular race that students have competed in, then temporal errors need not concern us. Description of the event includes a particular time and place and tomorrow is another day. If, on the other hand, they are presumed to indicate some skill or competency of the student, then they must also be presumed to have some constancy over time. Tomorrow is the same day in terms of traits and capacities and skills and understandings. At an ideological level the whole exercise depends on this. So if "skills" are developing then logically only the most recent performance should count. And if they are not developing then what are the students learning?

Table 1 Grade descriptions

Grade	core work	knowledge, competency	texts	wider reading	debates approaches	original and creative
pass 2 50-54	undertaken	adequate	basic		some familiarity	
pass 1 55-64	more	sound	sound		good general level of	

					familiarity	
credit 65-74	additional	sound	sound	done	apply a range	
distinction 75-84	considerable additional	advanced	advanced	considerable	broad familiarity and facility at applying	developing a capacity
high distinction 85-100	considerable additional	highest level	in depth	extensive	highest level of proficiency in applying	combining knowledge with
fail 0-49	fail to complete	fail to demonstrate				

Further to this, if they have actually learnt through the process of doing the project, or through any subsequent feedback, then the product becomes invalid because the state of the student is now different from that state when the product was produced, and another temporal error has been perpetrated.

In this sense, tests are premised on an assumption of student stasis; the more the student learns during or subsequent to any test information, the more that test information becomes outdated and hence in error.

Contextual errors

The grade descriptors do not mention context. But they imply a range of possible contexts, assessment modes, media, and processes. In order to make sense of such grades we must infer that the performances on which they are based are independent of the context in which they are produced; that is, they must represent a fixed measurable property of the student rather than a particular response to contextual events. It has been argued in this thesis that to believe this is an ontological error. Regardless, it is obvious that human behaviour, including cognitive behaviour, varies markedly according to context, so to reduce contextual error of the grades it would be necessary to specify the context of the events resulting in students' products, and the events resulting in the assessors product (the grade).

Without such contextual specification therefore the grades must of necessity be invalid.

Labelling errors

There are two labels; the label that describes the measure, and the label that describes what is measured. The assumption of these descriptors is that the measure can exist independent of what is measured. That grades have a reality independent of what is being graded. That administrative convenience can become a substantive reality. As indeed it will. But at what cost to professional integrity or student justice?

And even if this assumption is not nonsense, there is still the problem of the meaning of the grade. As I have indicated, the grade demarcations are so vague that errors within each criteria must be immense. Further, once the criteria become combined into a single dimension all information about individual criteria is lost, so all meaning related to the criteria likewise dissolves.

Attachment errors

As I have reiterated in many places in many ways in this thesis, information gained from tests is information about an event in which an individual student is an element. Any attempt to attach the description or data to the student, rather than to the total event, is an ontological slide. Attempts to not only attach to the student, but to some particular conceptual entity which the student is fantasised to have, takes us even deeper into the ontological bog. Error is reduced as the completeness of the event is recaptured. Such recapturing, of course, nullifies the use of simple numerical and graded categories.

In this case we have, in terms of the definitions of the grades, at least six independent classification events, all of which are supposed to contribute to the final grade. Error is indicated by any differences or confusions of grade within or among such events.

Frame of reference errors

The criteria would appear to indicate the Specific frame. Within each criteria there are indicators of grade demarcations. However, these are hardly adequate for specifying any standards. What is the difference between basic, sound, advanced, and in-depth? How do you draw fine lines between some familiarity, good general level of familiarity, broad familiarity and facility at applying? And how do you differentiate between developing a capacity for creativity, and combining knowledge with creativity? How else would you know a capacity was being developed than by relating it to knowledge? Obviously within the specific frame the indicators for cut-offs are hopelessly inadequate, and in this frame the system is grossly invalid.

Perhaps though this is unfair. Perhaps it is only political fashion that has forced this appearance of competency. The word "highest" appears twice, and this is obviously a normative term belonging to the General frame. Yet there are no percentiles given for grade boundaries, so standards are not possible to define within this frame. There are of course marks given that are appropriate to each grade. The Calender makes it clear, or at least implies strongly, that these marks are awarded as subdivisions of the grade, rather than that the grades are based on some previously determined marks. What is done in practice is moot. Regardless, the system is unworkable in the General frame, because there are no guidelines in this frame to decide grade boundaries. Within this frame therefore immense errors of miscategorisation must be expected.

In the Judge's frame, where as the reader will recall there is no error by definition, there is no problem. There never is. Judges have no problem differentiating between more core work, additional core work, and considerable additional core work. Even when, as appears to be the general case from the descriptions of courses given in the Calender, no core is specified. Or even, indeed, between the different "soundness" that differentiates pass level 1 from credit when applied to sound knowledge and competencies, and the sound understanding of texts.

It seems apparent that the criteria here are a competency smoke screen, a vague set of hints that allow assessors to continue to do what they have traditionally done; create a comparative order of merit of doubtful meaning , and at the same time allocate rather arbitrary grade boundaries to the rank order. The specification of criteria, naive and inadequate as they are, nevertheless fortifies the "scientism" of the Judge's frame, armouring its uncertain certainties with a coating of current assessment dogma.

Instrumental errors

With a plethora of assessment modes-assignments, practical work, observations, tests, examinations, it is sometimes difficult to actually locate the instrument, the "objective" machine that makes the measure. And of course there is no such objective machine. The fantasy that tests of various kinds are measuring instruments unfortunately remains a prevailing myth in the assessment of persons. The assessment modes are merely techniques used to fix a performance in time and space, to give it reality through some semblance of permanency. This allows, at least theoretically, independent judgments to be made of their "quality" or relative merit.

In practice the actual instrument, the place where the standard resides, the conceptual theory-practise link is established, the mark is produced, the comparisons are made, and the categorisations established -- all of these exist inside the mind of the examiner. So there is no objective instrument, and the assessment is clearly in the responsive mode, subject to all the normal variations and anomalies of idiosyncratic subjective judgments. Single examiners, which is the norm for university assessments, disguises this reality by nullifying in advance all competing judgments.

Categorisation errors

Within each criteria the categorisation boundaries are defined by words or phrases of extraordinary vagueness and imprecision, when it is remembered that this purports to be the official description of the categories that determine students' futures.

For example, assuming that the "core work" for a particular course has been precisely defined, then it might indeed be possible to determine whether it had been "undertaken." Or even if "more" than the required work was done, meriting a pass 1 classification. But how to distinguish this "more" from the "additional" core work required for a credit, or the "considerable additional" work required for a distinction or a high distinction, is unspecified. And how does the "sound" knowledge and competency required for a pass 1 differ from the "sound" knowledge and competency required for a credit, and in what way is that different from the "advanced" knowledge and competency required for a distinction or the "highest level" required for a high distinction? Surely it would be easier to be honest and say: "Rank order the students somehow and then draw arbitrary grade boundaries!"

Comparability errors

How are estimates for different criteria to be summated? The meaning of the final grade can only have a meaning in relation to the criteria if the loadings for each criteria are transparent, for how can we compare grades if they can mean different things. And how can we compare them anyway? How does "developing a capacity for original and creative work" in Commercial Law B compare with the same description in Human Resource Management or Mathematics 1A or Cognitive Science? What could "developing a capacity" possibly mean in any context, for that matter? And how can you compare the core work between subjects when it isn't specified in most cases? Indeed, if it isn't specified in some detail the whole grade description structure is entirely unworkable within a subject, for how could "additional" be judged without knowing what it was additional to?

Prediction errors

Whilst there are no overt predictions made in terms of these grades, there are some covert ones of immense significance. Certainly entry to higher degree programs is largely determined by the grades obtained, so there is an implicit prediction that students with lower grades are less suited to such further work. And, of course, students who fail are predicted as unsuited to qualify for work in particular fields.

Performance in academic course work, even if it could be accurately assessed, is very different from performance in professional work contexts. Yet the former is often, and increasingly, a necessary prerequisite for the latter. So the predictive validity of the grades would seem to be of vital importance, especially in those professions that require academic qualifications.

As indicated in Chapter 15, predictions about job performance on the basis of any selection criteria tend to be very low indeed, and correlations of 0.3 are considered very adequate. That this is ten percent better than pure chance indicates the immensity of the predictive error, and the extraordinary extent of the social injustice perpetrated through such mechanisms.

Logical type errors

Referral to Table 1 indicates there are six elements to the class of each grade. Are all elements required for the grade to be awarded? Or are five out of six enough? Or is one element enough for a higher grade? Could a person graded pass 2 be at a high distinction in five elements and be categorised pass because they had not done wider reading? How would we know that? If the elements must all be attained for a given level of grade then necessarily the lowest level in any element will alone determine the grade. If individual common sense gives the answers to these questions what can grades mean when common sense is so disparate?

Attention to possible logical type errors of this sort indicate inevitable massive confusion and thus error in the interpretation of these grades.

Value errors

What are some of the value errors implicit in this system? An obvious one is that "more and less" is synonymous with "better and worse." This shows very clearly in the descriptors for core work, knowledge and competency, and wider reading. The clear implication of these columns is that more is better.

This has considerable social as well as semantic significance. There is a value clearly implied that students should do more work than is specified or required, and that merit is accumulated through such activity. There are uncomfortable parallels here with current work practices in a competitive market, where workers are increasingly expected to work longer hours for no additional remuneration, and this exploitation becomes twisted by ideology to become a symptom of professionalism.

Another value, whose implications influence comparability errors, is that of terribly ordered learning. The six criteria must march along in unison otherwise they are unusable. It seems, for example, that original and creative thinking can only occur after masses of core, and additional conceptual work, has been understood. Is this true? Cannot innovative practical methodologies be constructed with very little specific knowledge? Cannot original and creative practical experiments and equipment design be produced to specifications with almost no knowledge of background theory? The limiting of the terms original and creative to the top two grades involves very prejudicial assumptions.

Consequential errors

How quickly and how intensely do students accept the judgments of their assessors as to the relative merit and idiosyncratic opinion (disguised as absolute value) of their academic performance? To what extent is the camouflage of error, the appearance of certainty, a predominant factor in this acceptance? To what extent does such acceptance affect later work, either positively or negatively? To what extent is the academic student constructed by the apparently objective measurements of their grades?

Such effects may be consistent within discernible sub-groups of students, or may be individually differentiated. Regardless, the questions indicate a particular category of invalidity, and in fairness to all students demand answers if the extent of invalidity for this criteria is to be explicated.

Not a problem

Does the confusion with its attendant error that is evident here create a problem for assessment in academia? It would seem not. Hopeless as the descriptors are, they are probably no better or worse than those they replaced, nor of others elsewhere. Academics just do not seem to problematise confusion and error in the measurement of "standards," at least not in academic discourse.

Is validity an issue? I checked the journal Assessment and Evaluation in Higher Education. Of a total of 195 articles in this journal from 1986 to 1996 only nine dealt, directly or by implication, with the problem of error, or inconsistency, or lack of validity in grading or marking. Of these nine there were three articles on validity which did not deal with inconsistency or error as any sort of a problem or issue. Four dealt with marker reliability, and two of these trivialised the notion of error in their conclusions.

Closer to home, Orrel's (1997) examination of the thinking-in-assessment of "everyday academics" revealed sometimes some angst in assigning a grade, but little concern that the "standard" itself might be illusory. And she commented that "A notable silence in the academic's discourse was any reference to the considerable technical measures that exist for assuring validity and reliability in assessment"(p397). But then, as they were clearly in the Judges frame of reference, such comment would have constituted a mind-shattering contradiction.

Conclusion

In the vernacular, it's a matter of "no worries, mate, business as usual!"

Part 7: Concluding statement

Chapter 20: Out of the fog

Chapter 20: Out of the fog

This study was begun to answer one fundamental question: How is error in measurement of standards obscured in most practical events involving assessment of persons?

Before I commenced work on this thesis I had already worked on this particular aspect for two years, and had written about ten chapters for a book on the subject. Further work during the past two years at Flinders University has developed and enlarged the scope of the work. As well, I have traversed some side roads, taken some wrong turnings, and come to a few dead ends. For example, at one stage it seemed the whole focus of the work would be on competencies. At another point interviews with assessment experts, administrators, teachers and students loomed large on the agenda. So at various times I was diverted from the main topic but always returned to it, often with fresh insights.

Tying the focus to the concepts of validity and invalidity was a relatively late development, only possible after the literature on validity was reframed as an advocacy for the test taker. The centrality of comparability to the whole assessment issue was similarly a late discovery.

I am personally pleased at the outcome. I can now make some sense out of what seemed non-sense; I have shown how some of the fudging was accomplished, and why it was important, in terms of social stability, to do so. At the same time I have, I believe, forged a powerful tool for the analysis of invalidity of assessment, and hence of error in the categorisation of individual persons--a tool based on a shift in positioning from test giver to test taker.

In a rational world the thirteen sources of invalidity, developed in many cases by reframing and repositioning the accepted scholarship in the field of assessment, should be sufficient to halt the conceptual blindness, the blatant suppression of error, the subtle fudges, and the myth of certainty that permeates the "science" and expertise of categorising people. Full acceptance and individual specification of even one of these sources could revolutionise current practice. However, as the study indicates, the world in which assessment resides is far from that rational world to which much of the writing in this thesis appeals.

I have tried to be clear about some of the forces that work on all of us that will encourage the reader to react strongly and negatively to many of my arguments, to dismiss them as anathema. The work is immoral in that it conceptually threatens the inviolability of standards and their measurement, a lynch pin of the cultural production of the modern individual. And it is revolutionary in that action based on its conclusions would destabilise to a point of destruction many, probably most, educational and work practices that result in the categorisation of people.

On the other hand, the basic contentions of this project are not contentious at the top levels of evaluation in Education, Medicine, or Law: Ph D theses in Education are assessed by different examiners and it is expected that such assessors will often differ in their judgments of quality; when expert opinions are sought in medicine both diagnosis and treatment prescriptions may differ markedly; and the seven judges in the high court often give conflicting verdicts.

The work could be criticised as being unduly negative. Even if the claims of the thesis are true, or partially true, is its position not destructively unhelpful? We need to categorise people, so take away the standard and what remains? How can people live with the certainty of uncertainty? At the very least, give us an alternative. And whilst I have not developed the alternatives, I have certainly presented them. The Responsive frame has many developed modes of assessment within its boundaries. The chapter on quality clearly indicates one way to go. We live in a world of complexity and uncertainty, a fuzzy multi-dimensional world of immense variety and diverse interpretations. What is challenged in this work is the myth that this complexity can be reduced to simple linear dimension by some sort of examination, as a preliminary to comparing with some standard of adequacy somewhere defined.

This thesis does not contend that people cannot be pinpointed along such dimensions, butterflies permanently fixed on the board. It happens to millions every day. What is shown is that such categorisations are inevitably permeated with confusion, uncertainty and error, that genuine rather than fudged estimates of much of this error can be made, and that this particular violation of the human mind and spirit will continue until they are.

References

(APA) American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington: American Psychological Association.

Apple, M. (1982). *Education and power*. Boston: Routledge and Kegan Paul.

Arendt, H. (1969). *On violence*. London: Penguin.

Australian National Training Authority. (1994). *Towards a skilled Australia: A national strategy for vocational education and training* : Australian National Training Authority.

Ayers, W. (1993). *To teach: The journey of a teacher*. New York: Teachers College Press.

Ball, S. (1994). *Education reform*. Buckingham: Open University Press.

Barone, T. (1992). Beyond theory and method: A case of critical storytelling. *Theory into Practice,* 31(2), 143-146.

Barton, L., Whitty, G., Miles, S., & Forlong, J. (1994). Teacher education and teacher professionalism in England: some emerging issues. *British Journal of Sociology in education, 15(4),* 529-543.

Bateson, G. (1972). *Steps to an ecology of mind*. New York: Ballantine Books.

Bateson, G. (1979). *Mind and nature*. London: Wildwood House.

Becker, H. (1990). Generalising from case studies. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education: the continuing debate*. New York: Teachers College, Columbia University.

Beevers, B. (1993). Competency-based training in TAFE: Rhetoric and reality. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*, . Canberra: Australian College of Education.

Behar, I. (1983). *Achievement Testing*. Beverly Hills: Sage Publications.

Beittel, K. (1984). Great swamp fires I have known: Competence and the hermeneutics of qualitative experiencing. In E. Short (Ed.), *Competence: Inquiries into its meaning and acquisition in educational settings*, (pp. 105-122). Lanham, MD: University Press of America.

Benett, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment and Evaluation in Higher Education, 18(2)*, 83-93.

Berk, R. (1986). A consumers guide to setting performance standards on criterion reference tests. *Review of Educational Research, 56(1)*, 137-172.

Biesta, G. (1994). Education as practical intersubjectivity: towards a critical-pragmatic understanding of education. Educational Theory, 44*(3)*, 299-317.

Bloom,B. (Ed.). (1956). *Taxonomy of educational objectives; Handbook 1, cognitive domain*. New York: David Mc Kay.

Bloom, B. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.

Bloom, B., Hastings, J., & Madaus, G. (1964). *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Borthwick, A. (1993). Key competencies - Uncovering the bridge between the general and vocational. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*, . Canberra: Australian College of Education.

Bourdieu, P., & Passeron, J. (1977). *Reproduction in education, society and culture*. London: SAGE Publications.

Bowden, J., & Masters, G. (1993). *Implications for higher education of a competency-based approach to education and training*. Canberra: Australian Government Publishing Service.

Bracht, G., & Glass, G. (1968). The external validity of experiments. *American Educational Research Journal, 5(4)*, 437-474.

Broadfoot, P. (Ed.). (1984). *Selection, certification and control*. London: The Falmer Press.

Brown, R. (1973). *Religion and violence*. Philadelphia: The Westminster Press.

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge: Harvard University Press.

Bucke, R. (1969). *Cosmic consciousness*. New York: E.P. Dutton Co.

Burchell, G., Gordon, C., & Miller, P. (Eds.). (1991). *The Foucault effect*. London: Harvester Wheatsheaf.

Burgess, R. (Ed.). (1985). *Issues in educational research: Qualitative methods*. London: The Falmer Press.

Burton, N. (1978). Societal standards. *Journal of Educational Measurement, 15(4),* 263-273.

Cairns, L. (1992). Competency-based education: Nostradamus's nostrum. *The Journal of Teaching Practice, 12(1)*, 1-32.

Camera, H. (1971). *Spiral of violence*. London: Sheed and Ward.

Campbell, J. (1956). *Hero with a thousand faces*. New York: Meridian Books.

Carr, W., & Kemmis, S. (1983). *Becoming critical*. Geelong: Deakin University Press.

Cherryholmes, C. (1988). *Power and criticism.* New York: Teachers College Press.

Cherryholmes, C. H. (1988). Construct validity and the discourses of research. *American Journal of Education, 96(3)*, 421-457.

Clough, E. E., Davis, P., & Sumner, R. (1984). *Assessing pupils: a study of policy and practice*. Windsor: NFER-Nelson.

Codd, J. (1985). *Curriculum discourse: text and context.* Paper presented at the National Conference of the Australian Curriculum Studies Association, La Trobe University, Melbourne.

Codd, J. (1988). The construction and deconstruction of educational policy documents. *Journal of Education Policy, 3(3),* 235-247.

Collins, C. (Ed.). (1993). *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.

Collins, R. (1979). *The credential society*. Orlando: Academic Press Inc.

Cox, R., & 1965. (1965). *Examinations and higher education: A survey of the literature*. London: Society for Research into Higher Education.

Cresswell, M. (1995). Technical and educational implications of using public examinations for selection to higher education. In T. Kellaghan (Ed.), *Admission to higher education*, . Dublin: Educational Research Centre.

Cronbach, L. (1969,). *Validation of educational measures.* Paper presented at the The 1969 invitational conference on testing problems: Towards a theory of achievement measurement.

Cronbach, L. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L., Rajaratman, N., & Gleser, G. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, XVI(2).*

Cronbach, L. J. (1990). *Essentials of psychological testing*. (Fifth ed.). New York: Harper and Row.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: performance assessment. *Educational Researcher, 23(5),* 11-18.

Docking, R. (1995, January 1995). Competency: What it means and how you know it has been achieved. *NTB Network- Special Conference Edition, 18.*

Donmeyer, R. (1990). Generalizability and the general case study. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education*, . New York: Teachers College, Colombia University.

Downs, C. (1995). *Key competencies: A useful agent for change?* . Richmond: National Centre for Competency Based Assessment and Training.

Eisner, E. (1988). The primacy of experience and the politics of method. *Educational Researcher, 17(3),* 15-20.

Eisner, E. (1990). The meaning of alternative paradigms. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: Sage Publications.

Eisner, E. (1991b). Taking a second look: Educational connoisseurship revisited. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*, (pp. 169-187). Chigago: The National Society for the Study of Education.

Eisner, E., & Peshkin, A. (Eds.). (1990). *Qualitative enquiry in education*.

Eisner, E. W. (1985). *The educational imagination*. (second ed.). New York: Macmillan.

Eisner, E. W. (1991). *The enlightened eye*. New York: Macmillan.

Fay, B. (1987). *Critical social science*. New York: Cornel University Press.

Feyerabend, P. (1988). *Against method*. London: Verso.

Finn, B. C. (1991). *Young people's participation in post-compulsory education and training* . Canberra: Australian Educational Council Review Committee.

Fish, S. (1980). *Is there a text in the class? The authority of interpretive communities*. Cambridge, Ma.: Harvard University Press.

Foucault, M. (1972). *The archaeology of knowledge*. London: Tavistock Publications.

Foucault, M. (1982a). Questions of method: an interview with Michel Foucault. *Ideology and Consciousness, 8(6)*, 3-14.

Foucault, M. (1982b). The subject and the power. In H. Dreyfus & P. Rabinow (Eds.), *Michel Foucault: Beyond structuralism and hermeneutics*, . Brighton: Harvester.

Foucault, M. (1988). *Politics, philosophy, culture: interviews and other writing*. New York: Routledge.

Foucault, M. (1992). *Discipline and punish*. London: Penguin.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18(9)*, 27-32.

Freud, S. (1963). *Civilisation and its discontents*. London: The Hogarth Press.

Friedenberg, E. (1969). , *Proceedings of the 1969 invitational conference on testing problems*, . Princeton: Educational Testing Service.

Garcia, G. E., & Pearson, P. D. (1994). Assessment and diversity, *Review of research in education*, (Vol. 20, pp. 337-391).

Garman, N. (1994). Qualitative enquiry: meaning and menace for educational researchers. In J. Smyth (Ed.), *Qualitative approaches in educational research*, (pp.

3-14). Adelaide: Flinders University of South Australia.

Garman, N., & Holland, P. (1995). the rhetoric of school reform reports: sacred, sceptical and cynical interpretations. In R. Ginsberg & D. Plank (Eds.), *Commissions, reports, reforms and educational policy*. Westport: Praeger.

Gillis, S., & Macpherson, C. (1995,). *Examination of the links between pre-employment qualifications and on the job competency based assessment.* Paper presented at the Australian Association for Research in Education, 25th Annual Conference, Hobart.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519-521.

Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement, 15(4)*, 237-261.

Golstein, H. (1979). Changing educational standards: A fruitless search. *Journal of the NAIEA,* 11(3), 18-19.

Gonzalez, E. J., & Beaton, A. E. (1994). The determination of cut scores for standards. In A. C. Tuijnman & T. N. Postlethwaite (Eds.), *Monitoring the standards of education*.

Good, F., & M, C. (1988). Grade awarding judgements in differential examinations. *British Educational Research Journal, 14(3)*, 263-281.

Green, M. (1994). Epistemology and Educational Research: the Influence of Recent approaches to Knowledge. *Review of Research in Education, 20*, 423-464.

Green, P. (1981). *The pursuit of inequality*. Oxford: Martin Robertson.

Guba, E. (1990). *The paradigm dialog*. Newbury Park: SAGE Publications.

Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Hacking, I. (1991). How should we do the history of statistics? In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect*, . London: Harvester Wheatsheaf.

Haertel E H. (1991). New forms of teacher assessment, *Review of Research in Education*, (Vol. 17, pp. 3-29).

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Hambleton, R., & Zaal, J. (Eds.). (1991). *Advances in educational and psychological testing*. Boston: Kluwer Academic Publishers.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing Company.

Hartog, P., & Rhodes, E. (1936). *The marks of examiners*. London: Macmillan and Co.

Harvey, L., & Greed, D. (1993). Defining quality. *Assessment and Evaluation in Higher Education, 18(1),* 9-34.

Horkheimer, M., & Adorno, T. (1972). *Dialectic of enlightenment*. New York: Herder and Herder.

House, E. (1991). Evaluation and social justice. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*, (pp. 233-247). Chicago: University of Chicago Press.

Howe, K. R. (1994). Standards, assessment, and equality of educational opportunity. *Educational Researcher, 23(8),* 27-33.

Hulin, C., Drasgow, F., & Parsons, C. (1983). *Item response theory: Application to psychological measurement*. Homewood, Illinios: Dow Jones-Irwin.

Huxley, A. (1950). *The perennial philosophy*. London: Chatto and Windus.

Illich, I. (1971). *Deschooling society*: Calder and Boyers Ltd.

Jackson, N. (1993). Competence: A game of smoke and mirrors? In C. Collins (Ed.), *Competencies: The competencies debate in Australian education and training, .* Canberra: Australian College of Education.

Jaeger, R., & Tittle, C. (Eds.). (1980). *Minimum competency achievement testing*. Berkeley: McCutchen Publishing Corporation.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement, Third edition*. New York: American Council on Education, Macmillan Publishing Company.

Johnston, B., & Dowdy, S. (1988). *Teaching and assessing in a negotiated curriculum.* Melbourne: Robert Anderson and Ass.

Johnston, B., & Pope, A. (1988). *Principles and practice of student assessment*. Adelaide: South Australian Education Department.

Jones, L. (1971). The nature of measurement. In R. Thorndike (Ed.), *Educational measurement: second edition,* (pp. 335-355). Washington: American Council on Education.

Kavan, R. (1985). *Love and freedom*. London: Grafton Books.

Keeney, B. (1983). *Aesthetics of change*. New York: The Guilford Press.

Kennedy, K., Marland, P., & Sturman, A. (1995). *Implementing national curriculum statements and profiles: corporate federalism in retreat.* Paper presented at the Annual Conference of the Australian Association for Research in Education, Hobart, 26-30 November.

Knight, B. (1992). Theoretical and practical approaches to evaluating the reliability and dependability of national curriculum test outcomes, : Unpublished article.

Korzybski, A. (1933). *Science and sanity*. Lakeville, Con: International non-Aristotelian Pub. Co.

Laing, R. (1967). *The politics of experience*. Harmondsworth: Penguin.

Lather, P. (1991). *Getting Smart: Feminist research and pedagogy with/in the postmodern*. New York: Routledge.

Lazarus, M. (1981). *Goodbye to excellence: A critical look at minimum competency testing*. Boulder: Westview Press.

LeCompte, M., Millroy, W., & Preissle, J. (1992). *The handbook of qualitative research in education*. San Diego: Academic Press Inc.

Levin, H. (1978). Educational performance standards: Image or substance*. Journal of Educational Measurement, 15(4)*, 309-319.

Lincoln, Y. (1990). The making of a constructivist. In E. Guba (Ed.), *The paradigm dialog*. Newbury Park: Sage Publications.

Lincoln, Y. (1995). Emerging criteria for quality in qualitative and interpretative research. *Qualitative Inquiry, 1*(3275-289).

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23(9),* 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20(8),* 15-21.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lorge, I. (1951). The fundamental nature of measurement. In E. Lindquist (Ed.), *Educational measurement,* (pp. 533-559). Washington: American Council on Education.

Madaus, G. F. (1986). Measurement specialists: Testing the faith - A reply to Mehrens. *Educational Measurement: Issues and Practice, 5(4),* 11-14.

Mager, R. (1962). *Preparing instructional objectives.* Palo Alto, CA: Feardon Publishers.

Marshall, C. (1990). Goodness criteria. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.

Masson, J. (1991). *Final analysis*. London: Harper Collins.

Masters, G. (1994, 17 March). *Setting and measuring performance standards for student achievement.* Paper presented at the Public Investment in School Education: Costs and Outcomes, Canberra.

Maturana, H., & Guiloff, G. (1980). The quest for the intelligence of intelligence. *Journal of Social Biological Structures, 3.*

Mayer, C. C. (1992). *Putting general education to work: the key competencies report* . Melbourne: Australian Educational Council and Ministers of Vocational Education, Employment and Training.

McDonald, R. (1994, October, 1994). Led astray by competence? Paper presented at the Australian National Training Authority, Brisbane.

McGovern, K. (1992). National competency standards - the role of the National Office of Overseas Skills Recognition. *The Journal of Teaching Practice, 12(1)*, 33-46.

Meadmore, D. (1993). The production of individuality through examination. *British Journal of Sociology in Education, 14(1),* 59-73.

Meadmore, D. (1995). Linking goals of governmentality with policies of assessment. *Assessment in Education, 2(1),* 9-22.

Melton, R. (1994). Competencies in perspective. *Educational Research, 36(3),* 285-294.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational Measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing Company.

Messick, S. (1989b). Meaning and values in test validation. *Educational Researcher, 18(2),* 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2),* 13-23.

Miller, A. (1983). *For your own good*. New York: Farrar, Straus, Giroux.

Miller, A. (1984). *Thou shalt not be aware*. London: Pluto Press.

Miller, C., & Parlett, M. (1974). *Up to the mark: a study of the examination game*. London: Society for Research into Higher Education.

Millman, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing company.

Mishler, E. (1986). *Research interviewing*. Cambridge: Harvard University Press.

Mitroff, I., & Sagasti, F. (1973). Epistemology as general systems theory: An approach to the design of complex decision-making experiments. *Philosophy of the social sciences(3)*, 117-134.

Moss, P. A. (1992). Shifting concepts of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62(3)*, 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23(2),* 5-12.

Mykhalovskiy, E. (1996). Reconsidering Table Talk: Critical thoughts on the relationship between sociology, autobiography and self-indulgence. *Qualitative Sociology, 19(1),* 131-151.

Nairn, A. (1980). *The reign of ETS* . Washington.

National Training Board. (1992). *National Competency Standards: Policy and Guidelines (Second Edition)* . Canberra: National Training board.

National Training Board. (1995, January 1995). Who's doing what? Assessment in Australia today. *NTB Network - Special Conference Edition,* 19-20.

Norris, N. (1991). The trouble with competence. *Cambridge Journal of Education, 21(3),* 331-341.

Nuttall, D. (1979). The myth of comparability. *Journal of the NAIEA, 11(3),* 16-18.

Nuttall, D., Backhouse, J., & Willmott, A. (1974). *Comparability of standards between subjects.* (Vol. 29). Oxford: Evans/Methuen Educational.

Oakley, A. (1991). Interviewing women. In H. Roberts (Ed.), *Doing feminist research*, . London: Routledge and Kegan Paul.

Orrell, J. (1996). Assessment in higher education: an examination of everyday academic's thinking-in-assessment, beliefs-about-assessment, and a comparison of assessment behaviours and beliefs. Unpublished Ph D, Flinders University of South Australia, Adelaide.

Partington, J. (1994). Double-marking students' work. *Assessment and Evaluation in Higher Education, 19(1),* 57-60.

Pawson, R. (1989). *A measure of measures*. London: Routledge.

Pearson, A. (1984). Competence: a normative analysis. In E. Short (Ed.), *Competence; Inquiries into its meaning and acquisition in educational settings,* (pp. 31-40). Lanham, MD: University Press of America.

Pennycuick, D., & Murphy, R. (1988). *The impact of graded tests*. London: The Falmer Press.

Perkins, D., & Salomon, G. (1988). Teaching for transfer. *Educational Leadership*(September), 22-32.

Persig, R. (1975). *Zen and the art of motorcycle maintenance: An enquiry into values*. New York: Bantam Press.

Persig, R. (1991). *Lila: An enquiry into morals*. London: Bantam Press.

Peters, M. (1996). *Poststructuralism, politics and education*. Westport: Bergin & Garvey.

Phillips, D. (1990). Subjectivity and objectivity: an objective enquiry. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education*. New York: Teachers college, Colombia University.

Popkewitz, T. (1984). *Paradigm and ideology in educational research*. London: The Falmer Press.

Porter, P., Rizvi, F., Knight, J., & Lingard, R. (1992). Competencies for a clever country: Building a house of cards? *Unicorn, 18(3),* 50-58.

Prigogine, I., & Stengers, I. (1985). *Order out of chaos*. London: Fontana.

Quine, W. (1953). *From a logical point of view*. New York: Harper and Row.

Rechter, B., & Wilson, N. (1968). Examining for university entrance in Australia: Current practices. *Quarterly Review of Australian Education, 2(2).*

Reilly, R., & Chao, G. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 33(1),* 1-55.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum and performance: A historical and comparative perspective. *Educational Researcher, 14(4)*, 5-20.

Rorty, R. (1991). *Objectivity, relativism, and truth*. Cambridge: Cambridge University Press.

Rose, N. (1990). *Governing the soul: The shaping of the private self.* London: Routledge.

Rosenberg. (1967). *On quality in art: criteria of excellence, past and present*. Princeton: Princeton University Press.

Royal Commission. (1974). *Report on the suspension of a high school student* . Adelaide: South Australian Government.

Sadler, D. R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education, 13(2)*, 191-209.

Sadler, R. (1995). Comparability of assessments, grades and qualifications. Paper presented at the AARE Conference, Hobart, 24 November.

Schmidt, F., Hunter, J., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: a red herring. *Journal of Applied Psychology, 66(2)*, 166-185.

Schnell, J. (1980). *The fate of the earth*. London: Picador.

Schwandt, T. (1990). Paths to enquiry in the social disciplines. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.

Scriven, M. (1991). *Evaluation thesaurus; fourth edition*. Newbury Park, Cal: SAGE Publications.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20(7),* 2-16.

Shepard, L. A. (1993). Evaluating test validity, *Review of research in education, 19.*

Sherman, R., & Webb, R. (1988). *Qualitative research in education*. New York: Falmer.

Slater, P. (1966). *Microcosm*. New York: John Wiley.

Smith, B. (1994). Addressing the delusion of relevance: Struggles in connecting educational research and social justice. In J. Smyth (Ed.), *Qualitative approaches in educational research*, (pp. 43-56). Adelaide: Flinders University of South Australia.

Smith, J. (1990). Alternative research paradigms and the problem of criteria. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.

Smith, J. (1993). *After the demise of empiricism: the problem of judging social and education inquiry*. Norwood, N.J.: Ablex Publishing Corporation.

Smyth, J. (Ed.). (1994). *Qualitative approaches in educational research*. Adelaide: Flinders University of South Australia.

Soucek, V. (1993). Is there a need to redress the balance between systems goals and lifeworld-oriented goals in public education in Australia? In C. Collins (Ed.), *Competencies: The competencies debate in Australian education and training*. Canberra: Australian college of Education.

Spearritt, D. (Ed.). (1980). *The improvement of measurement in education and psychology*. Hawthorne, Victoria: Australian Council for Educational Research.

Stake, R. (1991). The countenance of educational evaluation. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education: at quarter century*, (pp. 67-88). Chicago: the University of Chicago Press.

Stanley, G. (1993). The psychology of competency-based education. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.

Stern, D. (1991). *Diary of a baby*. London: Fontana.

Sternberg, R. (1990). T & T is an explosive combination: technology and testing. *Educational Psychologist, 25(3&4)*, 201-222.

Sydenham, P. (1979). *Measuring instruments: tools of knowledge and control*. London: Peter Peregrinus Ltd.

Taylor, C. (1994). Assessment for measurement of standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31(2),* 231-262.

Taylor, P. (1961). *Normative discourse*. Englewood Cliffs: Prentice-Hall, Inc.

The Flinders University of South Australia. (1997). *Calender*. Adelaide: Flinders University of South Australia.

The Social Development Group. (1979). *Developing the classroom group: How to make* *your class a better place to live in* . Adelaide: South Australian Education Department.

The Social Development Group. (1980). *How to make your classroom a better place to live in* . Adelaide: South Australian Education Department.

Thompson, P., & Pearce, P. (1990). *Testing times*. Adelaide: TAFE National Centre for Research and Development.

Thompson, W. (Ed.). (1987). *Gaia, a way of knowing*. Hudson: Lindisfarne Press.

Travers, E., & Allen, R. (1994). *Random sampling of student folios: a pilot study (10).* Brisbane: Board of Senior Secondary School Studies, Queensland.

Watzlewich, P. (1974). *Change*. New York: W Norton & Co.

Weiss, C. (1991). Evaluation research in the political context. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education; at quarter century*, (pp. 211-231). Chicago: The University of Chicago Press.

Wheeler, L. (1993). Reform of Australian vocational education and training: A competency-based system. In c. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.

Wiggins, G. (1988). Teaching to the (authentic) test. *Educational Leadership*(September), 41-47.

Wilbur, K. (1977). *The spectrum of consciousness*. Wheaton: Quest.

Wilbur, K. (1982). *Up from Eden: A transpersonal view of human evolution*. Boston: Shambhala.

Wilbur, K. (1991). *Grace and grit*. North Blackburn: Collins Dove.

Wilbur, K. (1995). *Sex, ecology, spirituality*. Boston: Shambhala.

Wiliam, D. (1995). Technical issues in criterion-referenced assessment: evidential and consequential bases. In T. Kellaghan (Ed.), *Admission to higher education*. Dublin: Educational Research Centre.

Williams, F. (Ed.). (1967). *Educational evaluation as feedback and guide*. Chigago: The National Society for the Study of Education.

Willmott, A. S., & Nuttall, D. L. (1975). *The reliability of examinations at 16+*. London: Macmillan Education Ltd.

Wilson, N. (1966). *A programmed course in physics, Form V*. Sydney: Angus and Robertson.

Wilson, N. (1969). Group discourse and test improvement. Unpublished data.

Wilson, N. (1969). A study of test-retest and of marker reliabilities of the 1966 commonwealth secondary scholarship examination. *ACER Information Bulletin, 50(1).*

Wilson, N. (1970). *Objective tests and mathematical learning*. Melbourne: Australian Council for Educational Research.

Wilson, N. (1972). *Assessment in the primary school*. Adelaide: South Australian Education Department.

Wilson, N. (1974). *A framework for assessment in the secondary school* . Adelaide: South Australian Education Department.

Wilson, N. (1985). *Young people's views of our world* (Peace Dossier 13). Melbourne: Victorian Association of Peace Studies.

Wilson, N. (1986). Programmes to reduce violence in schools. Adelaide: South Australian Education Department.

Wilson, N. (1992). *With the best of intentions*. Nairne: Noel Wilson.

Withers, G. (1995). Achieving comparability of school-based assessments in admissions procedures to higher education. In T. Kellaghan (Ed.), *Admission to higher education*. Dublin: Educational Research Centre.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment, *Review of research in education*, (Vol. 17, pp. 31-71).

Wolf, R. M. (1994). The validity and reliability of outcome measures. In A. C. Tuijnman & T. Neville Postlethwaite (Eds.), *Monitoring the standards of education.*

Wood, R. (1987). Aspects of the competence-performance distinction: Educational, psychological and measurement issues. *Curriculum Studies, 19(5),* 409-424.

Wood, R. (1987). *Measurement and assessment in education and psychology*. London: The Falmer Press.