Education Policy Analysis Archives

Volume 8 Number 16 **March 8, 2000** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

The Relationship between the Reliability and Cost of Performance Assessments

Jay Parkes
The University of New Mexico

Abstract
Performance assessments have come upon two major roadblocks: low reliability coefficients and high cost. Recent speculation has posited that the two are directly related such that cost must rise in order to increase reliability. This understanding may be an oversimplification of the relationship. Two empirical demonstrations are offered to show that more than one combination of sources of error may result in a desired generalizability coefficient and that it is possible to increase the number of observations while also decreasing cost.

The movement toward performance assessments for large-scale assessment purposes has encountered two major obstacles: first, such assessments have difficulty demonstrating highly reliable scores, and second, they tend to be very expensive. How these two problems are thought to be related influences the proposed solutions. This in turn will directly affect policies about the use of such assessments.

The problem of poor reliability in performance assessment scores stems from the lack of agreement among tasks, raters and other sources of measurement error. This is exhibited

in a variety of types of performance assessments by several concurrent lines of inquiry, including: those by Shavelson and colleagues (e.g. Shavelson and Baxter,1992; and Shavelson, Baxter, and Gao, 1993); those from the Vermont Portfolio Assessment program (e.g. Koretz, Klein, McCaffrey, and Stecher, 1994; Koretz, Stecher, Klein, and McCaffrey, 1994; and Koretz, Stecher, Klein, McCaffrey, and Deibert, 1994); and one by McWilliam and Ware (1994).

Shavelson and colleagues have worked primarily with performance assessments in elementary level general science. By using the framework of generalizability theory, they have demonstrated that the greatest contributing facet to low generalizability coefficients is the task (e.g. Shavelson, Baxter and Gao, 1993). Furthermore, they project that by increasing the number of tasks a higher generalizability coefficient will result. Koretz and colleagues have worked with portfolio assessments of math and writing and identified raters and tasks as sources of error variance (Koretz, Stecher, Klein, McCaffrey, and Deibert, 1994). They, too, explore the possibility of increasing the number of tasks and the number of raters to achieve a more acceptable estimate of reliability. McWilliam and Ware (1994) examined the assessment of young children's engagement, and identified the number of sessions or observations as being a large source of error variance. They estimated the minimum number of sessions that would be necessary to create an acceptably reliable assessment.

A second major concern with performance assessments is their high cost (Picus, 1994). Performance assessments are widely believed to be more expensive than multiple-choice testing (Catterall & Winters, 1994; Hardy, 1996; Linn, Baker & Dunbar, 1991; U.S. General Accounting Office, 1993), though the costs of performance assessments will vary considerably based on the exact nature of the assessment (Monk, 1996; U.S. General Accounting Office, 1993). Reckase (1995) demonstrated that it is possible to produce a writing portfolio assessment procedure that meets current standards of psychometric quality; but such a procedure, compared to current multiple-choice methods, would be a "very expensive alternative (p. 14)." White (1986), however, holds that, when designed properly, a direct assessment of writing can be conducted with comparable expense to that of multiple-choice assessment. This divergence notwithstanding, White (1986) recognized that the expenses are different for the two forms, the money being used mostly for raters in a direct assessment of writing. Hoover and Bray (1995) to some extent validated this claim by showing that the Iowa Writing Test could be conducted for approximately the same cost as the Iowa test of Basic Skills, albeit the former covered a much smaller domain than the latter.

These two problems of low reliability coefficients and high cost in performance assessment are often directly linked. If the solution to low generalizability is to increase the number of tasks, raters, etc., then the cost must also increase (e.g. Picus, 1994). There are a number of issues, however, that make this more complicated than it first appears.

The first issue is the automatic acceptance of the direct relationship between the number of observations in an assessment and the reliability of scores from that assessment. This acceptance is promulgated by a long history with the Spearman- Brown Prophecy Formula used to address this issue with objective item assessments. In a multiple-choice test, it is possible to estimate the number of items necessary to reach a desired reliability coefficient. For example, if a test contains 50 multiple-choice items and the reliability coefficient for scores from that test is 0.76, the Spearman-Brown Prophecy Formula can be employed to estimate how many items would need to be added to increase the reliability estimate to 0.85. There is direct (though asymptotic) relationship between the number of items used and the magnitude of the reliability coefficient. In a performance assessment, however, the relationship between a reliability estimate and the number of observations is more complicated because there are more sources of error. In a multiple-choice test, the

items represent the only source of error. In a performance assessment, tasks, raters, occasions and potentially many other sources of error are possible. The implication is twofold. First, there may be more than one combination of raters, tasks, etc. that will result in a reliability estimate of a given magnitude. Second, it is possible that fewer observations could lead to a larger estimate of the reliability of scores from a performance assessment. Therefore, it is no longer axiomatic that increasing reliability means adding more observations.

The second issue is that cost and reliability are seldom addressed simultaneously. By and large this is due to the methodologies employed for such projections. In an assessment procedure with multiple sources of error, the most common projective technique is a liberalization of the Spearman-Brown Prophecy Formula, the decision study, or d-study from the generalizability theory framework. The d-study approach to addressing the joint issues of cost and reliability is less than desirable in a couple of ways.

D-studies are often done one at a time by considering different combinations of sources of error. That means that when the first combination to reach the desired reliability estimate is reached, the process stops. If there are several combinations of sources of error that would satisfy the desired reliability threshold, they probably would not be uncovered in this manner.

The d-study approach does not take cost information into consideration, which leaves the direct relationship between the number of observations and cost to dictate the best combination of sources of error. Assuming that d-studies are conducted in such a manner that multiple combinations of sources of error are identified, all meeting a minimum reliability estimate, the one with the fewest total observations is likely to be selected for implementation. It might be possible that more total observations could actually be less expensive. Without explicitly examining cost information, there is no way to know for sure.

The goal should be an optimal assessment design where optimal is defined as the most reliable and least expensive. There is a technique that allows all of these issues to be handled simultaneously in one analysis. Sanders, Theunissen, and Baas (1989, 1991, 1992) proposed the use of a branch-and-bound integer programming algorithm which searches for and identifies the optimal number of levels for each facet while taking into account each facet's contribution to the generalizability coefficient and each facet's cost as well as any other practical constraint. This technique appears to be promising. It can exhaustively search all possible combinations of levels of facets, within given parameters, something that could be a daunting task to perform "by hand" using only psychometric constraints. Thus it gives reasonable assurance that the optimal solution has been located.

A second advantage of this technique is that it can accommodate a wide variety of logistical, economic, or other constraints. So cost data and reliability data, as well as other relevant issues, can be used simultaneously to define an optimal assessment design.

These issues and procedures will now be demonstrated using two different studies. The first study concludes that, depending on the definition of "optimal," there are many possible best combinations of facets to produce a predetermined generalizability coefficient. The second study produces data supporting the Sanders, et al. (1991) statement that it is possible to decrease the number of observations and/ or the total cost while increasing the generalizability coefficient. Both studies are based on the same set of data.

The Optimization Studies

Subjects. Fifty subjects enrolled in an undergraduate educational psychology class participated in the study. Twenty- eight percent of the sample were males and seventy-two percent were females. The sample also contained a mix of White, Asian- American, and

Hispanic subjects. By class, the sample consisted of freshmen (20%), sophomores (52%), juniors (21%), seniors (5%), with the remainder unidentified. The sample had taken an average of 1.26 writing courses with a range from 0 to 3.

Procedures. Each subject read three articles—one about instructional approaches, and two articles about performance assessments—prior to attending the first of two 2 1/2 hour sessions. During the first session, subjects filled out a demographic questionnaire and wrote a separate 300 to 500 word essay about each of two prompts. During the second session, subjects wrote the other two prompts. In total, they wrote an expressive piece and a persuasive piece about the instructional approaches and an expressive piece and a persuasive piece about performance assessments. Four different orders of the prompts were counterbalanced to allow investigation of practice effects or other effects that may arise by writing the essays in a particular order.

Scoring the essays. Three graduate students in Educational Psychology served as raters and were trained. These raters were given the scoring rubric and discussed it; then, they scored a sample paper as a group. Using a slightly modified version of the Diederich scale (Diederich, 1974), each rater then read all 200 pieces of writing. The seven items on the scale were summed to achieve each subject's score on each piece of writing.

The Variance Models

The studies are based on a three-facet mixed design: mode of discourse (m), writing prompt (p), and rater (r). The object of measurement is student's overall writing ability (s). In the data collection design, prompts are nested within mode (i.e., p:m) and both cross raters and students. In the generalizability framework, the variance model is:

$$\sigma^2_{(xsrpm)} = \sigma^2_s + \sigma^2_r + \sigma^2_m + \sigma^2_{p:m} + \sigma^2_{sr} + \sigma^2_{sm} + \sigma^2_{mr} + \sigma^2_{srm} + \sigma^2_{(p:m)s} + \sigma^2_{(p:m)r} + \sigma^2_{(p:m)sr} \quad (1)$$

The variance components for the sample in this study were estimated using the GENOVA software program (Crick and Brennan, 1983). Based on a review of the literature on modes of discourse (Crusius, 1989), there are at most five modes in existence. Therefore, for the estimation of variance components, the universe of modes was defined as having 5 levels. For all other facets, the universes were defined as infinite. The variance components estimated are shown in Table 1.

Table 1
Estimated Variance Components for Studies One and Two

Source of variation	Variance components
Subject (s)	5.8275728
Mode (m)	0*
Prompt:mode (p:m)	0*
Rater (r)	5.6756912
sm	0*
s(p:m)	2.6025238

sr	0.6714422
smr	0.3008503
sr(p:m)	11.8791415

*Note. Negative variance components were set equal to zero, following Brennan (1992).

For all subsequent optimization analyses, the relative model of measurement was used wherein the relative error variances were estimated through:

$$\sigma^2_{(\delta)} = \frac{\sigma^2_{sr}}{n_r} + \frac{\sigma^2_{sm}}{n_m} + \frac{\sigma^2_{srm}}{n_r n_m} + \frac{\sigma^2_{(p:m)s}}{n_m n_p} + \frac{\sigma^2_{(p:m)sr}}{n_r n_m n_p} \quad (2)$$

where n_r, n_m, and n_p are the number of raters, modes, and prompts respectively.

The G-coefficient of interest was therefore:

$$E\rho^2 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{(\delta)}} \quad (3)$$

Study One

In this study, results of a generalizability study and data describing the number of person-hours necessary to score the assessment have been used. Four different scenarios are presented, each with a different set of constraints, each producing a different optimal solution. The first scenario optimized the problem using only psychometric constraints; the second took a relative human factor constraint into consideration; the third used a specific human factor constraint; and the fourth used specific economic constraints.

The Optimization Scenarios

A branch-and-bound integer programming algorithm, a linear programming technique, was employed to estimate the optimal combination of raters, prompts within modes, and modes themselves. This investigation used the solver function of Microsoft EXCEL, version 5.0, to execute the algorithm. For all four scenarios, the variance components from Table 1 were entered into the worksheet. All four scenarios investigated shared a common objective function and a common set of constraints. In Scenarios 2, 3, and 4, additional constraints were considered. The common problem to be solved across all scenarios is:

Objective Function:

$$\text{Minimize } L = n_m\, n_{p:m}\, n_r\, ; \quad \textbf{(4)}$$

Subject to:

$$E\rho^2 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{(\delta)}^2} \geq 0.8 \qquad (5)$$

$n_m <$ or = 5, **(6)**

n_m $n_{p:m}$ n_r are integers, **(7)**

and n_m, $n_{p:m}$, and $n_r >$ or = 1. **(8)**

The objective function is to minimize the total number of observations needed. Constraint (5) specifies the minimal acceptable level of a generalizability coefficient. Constraint (6) specifies that there are no more than 5 possible modes of discourse. Constraints (7) and (8) ensure that solutions will be positive whole numbers.

In Scenario 1, the objective function defined in (4) subject to constraints (5) through (8) was submitted to the branch-and- bound search algorithm. The results of this search can be found in Table 2, which shows that, to attain a g-coefficient of at least 0.8, the minimum numbers are 4 modes with 2 prompts each while employing two raters to score each prompt in each mode. Based on data obtained from the sample, the average time needed to rate each prompt in each mode in this study was 0.092 hour (approximately 5.5 minutes). The total amount of time needed to rate the writings from ns subjects under any given scenario is then:

Total person-hours = n_m $n_{p:m}$ n_r n_s(.092) **(9)**

Applying Equation (9), the total person-hours needed for Scenario 1 for 50 subjects is 73.6.

Table 2
Results of Study One
Number of Cases Needed to Meet the Constraints

	Actual	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
Additional Constraints			$n_r \geq n_m(n_{p:m})$	$n_m(n_{p:m}) \leq 6$	$n_m(n_{p:m})n_r(50)(.092) \leq$ 60 person-hours $n_m(n_{p:m})n_r(50)(.092) \leq$ 70 person-hours
Mode	2	4	1	1	4
Prompt: Mode	2	2	4	6	2
Rater	3	2	5	3	2
Obj. Function	12	16	20	18	16
Manhours	55.2	73.6	92	82.8	73.6
G Coefficient	0.75	0.80	0.80	0.80	0.80

An apparent practical problem with Scenario 1 is the demand on the examinee. A better solution might be one in which the burden of reliability is shifted away from the demand on the examinee to a demand on ratings per piece of writing. In Scenario 2, a new constraint was added to shift this demand to ratings. The additional constraint and the results can be found in table 2. To attain a g-coefficient of at least 0.8 while minimizing the burden on the examinee, the minimal design is one in which each examinee responds to 4 different prompts in a single mode of discourse. Each piece of writing needs to be rated by 5 raters. Under this scenario, the total number of writings from each examinee is only four. However, the total amount of person-hours needed for the rating of 50 subjects increases to 92 person-hours.

In Scenario 3, a compromise between Scenarios 1 and 2 was investigated by constraining the total number of pieces to six or less (see table 2). Under this scenario, each examinee must produce 6 pieces of writing in a single mode. On the other hand, only 3 raters are needed for each piece to attain a g-coefficient of 0.8 or higher. The total person-hours for 50 subjects in this case is 82.8.

Scenario 4 investigated the cost factor. The lowest number of person-hours so far has been 73.6 in Scenario 1. Scenario 4 attempted to explore the possibility of a person-hour estimate lower than that. Table 2 illustrates the two constraints attempted, neither of which produced a feasible solution. In other words, it is not possible to expend less than 70 person-hours of rating activities to rate the writings used in this study for 50 subjects and still maintain a minimum g-coefficient of 0.8.

Conclusions from Study One

In a single-facet measurement situation, a multiple-choice exam for example, there is only one source of error to draw on to increase a reliability coefficient: items. So a one-to-one relationship exists between the number of the facet and the reliability coefficient: as the number of items increases, so does the reliability coefficient, albeit the relationship is asymptotic at some point. Also, there is a unique minimum number of items that will satisfy the desired reliability coefficient. For example, if a 50-item exam has a reliability coefficient of 0.69, the Spearman-Brown Prophecy Formula may indicate that in order to achieve a coefficient of 0.90, 83 items are needed. In a multi-faceted situation like the one represented here, the relationships are not so clear. With multiple facets, each contributing unequally in proportion to the size of its variance component to the generalizability coefficient, there is no simple one-to-one relationship. Scenario 1 uses psychometric constraints alone (as the Spearman-Brown Prophecy Formula or other projective techniques would) yet mode changes by 2 units, prompt within mode does not change, and raters decreases by one unit. Thus, in multi-faceted situations using only psychometric criteria, the relationship between the facets and the generalizability coefficient is not straightforward or simple.

Neither in a multi-faceted situation is there one combination which will uniquely fulfill the predetermined generalizability coefficient. The first step is to define optimal in some way. The optimization procedure allows a great deal of latitude in doing so. The four scenarios taken together demonstrate that there are many optimal combinations that will fulfill the predetermined generalizability coefficient.

Study Two

The second study is similar to the first except that instead of using person-hours as the

economic constraint, it employs dollar figures. Second, instead of minimizing the total number of observations in order to constrain costs, it uses total cost as the objective function. The variance model in Study Two is the same as that in Study One.

The Cost Data

The cost data for this study are taken from Hoover and Bray (1995), who report on cost information for an administration of the Iowa Writing Assessment. The assessment tested the writing skills of 30,000 school students from grades three to twelve, each of whom wrote two pieces of writing. Each sample was scored twice holistically and twice analytically. For this assessment, Hoover and Bray estimate that $138,000 was spent in developing the 40 writing prompts; $174,410 was spent to score the prompts; and $30,000 was spent for materials. This breakdown is consistent with a framework for examining costs explained by Hardy (1996). In order to use this information in the optimization procedure, base units of development, scoring and materials need to be developed. That is, figures need to be obtained that indicate how much adding one rating (for example) to the scenario will change scoring costs, or how much adding one prompt will change development and scoring costs. The cost of development hinges on the total number of prompts developed—in Hoover and Bray (1995), 40—therefore, each prompt costs $3450 to develop ($138,000/40). In that study, each examinee wrote two prompts. Had each written only one prompt, presumably only 20 prompts would have been developed. Therefore, the $3450 is divided by 2, the number of prompts each examinee responded to, producing a cost per prompt required of an examinee of $1725. So that represents the base unit cost for development. Therefore, the development cost function is

$$\$1725 n_{p:m} n_m,$$

where $n_{p:m}$ is the number of prompts each person must write per mode
and n_m is the number of modes.

To obtain the base unit cost for scoring, the total scoring cost ($174, 410) was divided by the number of subjects (30,000), the number of pieces per subject (2), and the number of raters or readings per piece (2) to produce a unit scoring cost of $1.43 per piece, per rater, per subject. The materials were estimated to cost $1.00 per subject. For the purposes of these analyses, the number of subjects was held constant at 50. Therefore, the total cost function, combining development, scoring and material costs, is:

$$\text{Total Cost} = \$1725 n_{p:m} n_m + \$1.43 n_{p:m} n_m n_r n_s + \$1.00 n_s \qquad \textbf{(10)}$$

The Optimization Problem

The variance components from Table 1, the cost function given in equation (10), and the number of prompts within modes, modes, raters, and subjects were entered into the EXCEL worksheet, and the following optimization problem was submitted for analysis.

Objective Function:

$$\text{Minimize } L = \text{Total Cost} = \$1725 n_p + \$1.43 n_p n_r n_s + \$1.00 n_s, \qquad \textbf{(11)}$$

subject to constraints (5) through (8) given in Study One.

The results are given in Table 3. Since the procedure was minimizing cost not the number of observation points, the optimal design includes more observation points (27 versus 12) but at less cost and a higher generalizability coefficient.

Table 3
Results of Study Two
Number of Cases Needed to Meet the Constraints

	Actual	**Optimal**
Mode	2	1
Prompt:Mode	2	3
Rater	3	9
Obj. Function	12	27
Total Cost	$7808	$7156
G Coefficient	0.75	0.80

Conclusions from Study Two

This second study provides empirical support for the claim made by Sanders, Theunissen, and Baas (1989) that it is possible to decrease cost while increasing the generalizability coefficient even when the total number of observation points increases.

Discussion

These studies serve as illustrations of the issues raised in the introduction. The first study demonstrates that it is possible to have many combinations of facets in an assessment design meet some predetermined level of reliability coefficient. The second study demonstrates the advantages of simultaneously considering cost and reliability data in the same analysis, namely, that it is possible to achieve a more reliable but less costly design.

Both of these points need to be taken in consideration during discussions about the cost implications of various solutions to the low reliability problem associated with performance assessment scores. If we assume that the only way to increase the reliability is to increase the number of observations and/ or we assume that increasing reliability will automatically increase cost, these stumbling blocks will not be removed. Policy makers will continue to be very reluctant to choose performance assessments as parts of their assessment plans.

These demonstrations represent a narrow perspective though and were designed to demonstrate only the two issues already mentioned. They are narrow in two ways. First, they may oversimplify the estimation of true costs of performance assessments. Second, they address only reliability and cost and not other concerns.

The costs associated here with performance assessments are expressed in dollars and cents and are rather simple. For example, development costs would change depending on the number of examinees (Parkes, 1996). More examinees would require that more prompts be developed and the cost would probably change in some exponential fashion. This relationship is held constant by assuming the same number of examinees in each scenario. There are also many other ways to conceptualize cost, some of which would be very difficult

to quantify. Monk (1996) and Picus (1994) describe the difficulties in determine the actual "costs" of a performance assessment. There are, of course, the financial expenditures associated with an assessment system. But more nebulously, there will be expenditure of time by students, teachers, and administrators to conduct these assessments. There is also cost in terms of what curriculum changes are made to accommodate the testing. That is, what would students be learning in the time taken for assessment.

The studies reported here are also narrow in that they address only reliability and cost and not other concerns. And there are plenty of other considerations that are equally as important or more important in the design of a performance assessment besides reliability and cost. The content sampling issue is one of these. Deciding how many tasks should constitute an assessment should probably be addressed in terms of content coverage first. Though certain constraints could be added to an optimization problem to account for content coverage issues, it probably not best to handle the issue in that manner. This approach treats each facet of the design equally or weights it based on its contribution to error variance. It therefore works on the implicit assumption that one rater means essentially the same thing as one task, which means essentially the same thing as one occasion, etc. But raters and tasks and occasions all serve different purposes in the assessment and contribute different things to the construct validity of the scores. So to trade three tasks for five ratings is, at best, contrived.

These issues provide a necessary context for the studies reported here but should not distract attention from the two central issues of this paper. First, more than one combination of sources of error may result in a desired generalizability coefficient. Second, it is possible to increase the number of observations while also decreasing cost.

Conclusion

The notion that only one design will generate a g-coefficient of a given value is not accurate. There are many possible combinations of facets, depending on how the optimal solution is defined, that will meet a desired g-coefficient value. The relationship between an assessment design and a corresponding generalizability coefficient needs to be more broadly understood.

The inference that generalizability coefficients and the number of observations are directly related is inappropriate. It is possible that several different designs would achieve acceptable generalizability coefficients. Similarly, a direct relationship between cost and reliability is not exact. Study Two shows that it is possible to increase the generalizability coefficient and the number of observations while decreasing the total cost of the assessment.

The bottom line for policymakers and those involved in performance assessment programs is that it is theoretically possible to have both a reliable and cost-effective performance assessment system. Assuming that low cost is the "line in the sand," those developing performance assessments should not assume that means they must minimize the number of ratings or the number of pieces in an assessment. Indeed, increasing certain aspects, like ratings, might actually end up being cheaper and still produce more reliable scores.

References

Catterall, J. S. & Winters, L. (1994, August). Economic analysis of testing: Competency, certification, and "authentic" assessments (CSE Technical Report 383). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Crick, J. E., and Brennan, R. L. (1982). GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual). Iowa City: ACT.

Crusius, T. W. (1989). *Discourse: A critique and synthesis of major theories*. New York: Modern Language Association of America.

Diederich, P. B. (1974). *Measuring Growth in English*. Urbana, IL: National Council of Teachers of English.

Hardy, R. (1996). Performance assessment: Examining the costs. In M. B. Kane & R. Mitchell *Implementing perfomance assessment: Promises, problems, and challenges* (pp. 107-117). Mahwah, NJ: Lawrence Erlbaum Associates.

Hoover, H. D., and Bray, G. (1995, April). The research and development phase: Can a performance assessment be cost effective? Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Koretz, D., Klein, S., McCaffrey, D., and Stecher, B. (1994). Interim report: The reliability of the Vermont portfolio scores in the 1992-93 school year. (RAND/ RP - 260). Santa Monica, CA: RAND.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13* (3), 5-16.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., and Deibert, E. (1994). Can portfolios assess student performance and influence instruction?. (RAND/ RP-259). Santa Monica, CA: RAND.

Linn, R. L., Baker, E. L., and Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Monk, D. H. (1996). Conceptualizing the costs of large-scale pupil performance assessment. In M. B. Kane & R. Mitchell *Implementing perfomance assessment: Promises, problems, and challenges* (pp. 119-137). Mahwah, NJ: Lawrence Erlbaum Associates.

McWilliam, R. A., and Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18* (1), 34-47.

Parkes, J. (1996, April). Optimal designs for performance assessments: The subject factor. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Picus, L. O. (1994, August). A conceptual framework for analyzing the costs of alternative assessment (CSE Technical Report 384). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*(1), 12-14.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54*, 587-598.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1991). Maximizing the coefficient of generalizability under the constraint of limited resources. *Psychometrika, 56*(1), 87-96.

Sanders, P. F., Theunissen, T. J. J. M., and Baas, S. M. (1992). The optimization of decision studies. In M. Wilson (Ed.) *Objective Measurement: Theory into Practice (Vol. 1).* Norwood, NJ: Ablex.

Shavelson, R. J. and Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49*(8), 20-25.

Shavelson, R. J., Baxter, G. P., and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

U.S. General Accounting Office. (1993). Student testing: Current extent and expenditures, with cost estimates for a national examination (GAO/PEMD-93-8). Washington DC: U.S. General Accounting Office, Program Evaluation and Methodology Division.

White, E. M. (1986). Pitfalls in the testing of writing. In K. L. Greenberg, H. S. Wiener, and R. A. Donovan (Eds.), *Writing Assessment: Issues and strategies*. New York: Longman.

About the Author

Jay Parkes

Educational Psychology Program
128 Simpson Hall
University of New Mexico
Albuquerque, NM 87131-1246

E-mail: parkes@unm.edu

Jay Parkes is an Assistant Professor of Educational Psychology at the University of New Mexico. He received his Ph.D. in Educational Psychology from the Pennsylvania State University in 1998 specializing in applied measurement and statistics. His research interests include generalizability in performance assessment and cognitive and motivational aspects of performance assessments.

---

casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board