Validation Review Instrument

Based on the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), the work of Kane (2013) and Messick (1995), and other scholars in the field of validity, this instrument is designed to support article reviews with an eye toward issues of validity in general, and specifically toward the validation of a particular interpretation and use of scores.

The sections and items in this instrument follow the framework shown in Figure 1. Simply stated, each of the wedges represent different sources of evidence that either support or challenge the specific interpretation / use argument (IUA) being validated. The IUA relies on the quality of the scores that inform it, and thus, the five sources of evidence are pictured supporting the IUA through the measurement and methodology issues on which the IUA rests. For more information, see the description of the validity framework in the research team's shared folder (link removed for blind review).

Figure 1. Validity Framework

0.0) Information about the Article Reviewed

Information in this section relates to the article and its intended purpose without regard to whether that purpose is directly or indirectly related to the interpretation / use argument (IUA) under investigation.

0.1) Reviewer
Which member of the research team is completing this entry?

- ○ Reviewer 1 (blinded for peer review)
- ○ Reviewer 2 (blinded for peer review)
- ○ Reviewer 3 (blinded for peer review)
- ○ Reviewer 4 (blinded for peer review)

0.2) First Author
Copy and paste the last name of the first author here.

0.3) Year of Publication
Copy and paste or type the four-digit year of publication here.

0.4) Full Title
Copy and paste full title or the publication reviewed here up to the colon (:). Do not include the subtitle.

0.5) Type of Publication
What type of publication is the reviewed piece?

- ☐ (A) Report of Empirical Research
- ☐ (B) Conceptual / Theoretical Piece
- ☐ (C) Systematic Review or Meta-Analysis
- ☐ (D) Program Evaluation Report
- ☐ (E) Other Technical Report
- ☐ (F) Methodological Piece
- ☐ Other

0.6) Purpose of the Study

What is the original purpose of the study reviewed (not its purpose as related to the IUA being validated)?

This should be 1) copied and pasted from the abstract, 2) copied and pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

0.7) Methodological Approach

What general methods were used in the study reviewed?

This should be 1) copied and pasted from the abstract, 2) copied and pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

0.8) General Findings

What are the findings reported in the paper (not necessarily the findings that relate to the IUA being validated in this review)?

This should be 1) copied and pasted from the abstract, 2) copied and pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

1.0) Interpretation / Use Argument (IUA)

"To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the test scores. Validation therefore requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores. Public claims require public justification" (Kane, 2013, p.1).

The Interpretation / Use Argument (IUA) "includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use)" (Kane, 2013, p. 2).

"The kinds of evidence required for validation are determined by the claims being made, and more-ambitious claims require more evidence than less-ambitious claims" (Kane, 2013, p. 3).

"More-ambitious interpretations (e.g., a construct interpretation or a causal claim) tend to be more useful than less-ambitious interpretations, but they are harder to validate" (Kane, 2013, p. 37).

This review will address how the reviewed article contributes to the following IUA:

STUDENT SCORES ON LARGE-SCALE STANDARDIZED TESTS CAN AND SHOULD BE USED AS A VALID MEASURE OF TEACHER EFFECTIVENESS

For the purposes of this review, a **standardized test** will be defined as a test of predetermined content and/or standards that meets the following criteria:

1. The test is administered using standardized procedures and materials designed to ensure similar testing conditions across all test takers
2. All test takers respond to the same test items or a selection of items drawn from an approved bank of items
3. The test scores are calculated and/or reported in a manner designed to support comparisons between individual test takers or groups of test takers
4. The test is administered on a large scale (e.g., nationally, in one or more states, or throughout a large district).

The **score that informs the IUA** will be defined as *an estimate of teacher effectiveness* derived from student scores on one or more administrations of a standardized test. These teacher effectiveness scores are generated in such a way as to support comparisons between individual teachers or groups of teachers.

The earliest literature on validity began with tests, in the the most narrow and literal sense of the word. As a result, the language of the validity literature still reflects this heritage. Since the scores which inform the IUA under investigation are estimates of teacher effectiveness that have been derived from student standardized test scores, then the **"test"** in question must be defined as *the analytical procedure* used to derive teacher effectiveness estimates, and not the standardized tests that the students take. This distinction will be discussed in more detail in explanatory notes added to each section of the instrument where it is relevant.

Note: It may be helpful to complete the items in this section last.

1.1) Position on IUA being Validated

What kind of position does the publication reviewed take on the IUA above?

NOTE: This is the author's take on the IUA as revealed through the publication reviewed, not the reviewer's assessment of how well the publication supports the IUA in question.

- ◯ SUPPORTIVE (*The publication proposes or supports the IUA*)
- ◯ SOMEWHAT SUPPORTIVE (*The publication is generally supportive of the IUA, though it may advise some caution.*)
- ◯ NEITHER SUPPORTIVE NOR CHALLENGING (*The publication provides evidence or discussion in support of the IUA as well as evidence or discussion which challenges it.*)
- ◯ SOMEWHAT CHALENGING (*The publication critiques, challenges, or offers caution about the IUA while acknowledging its appropriate use.*)
- ◯ CHALENGING (*The publication challenges, opposes, or argues against the IUA.*)
- ◯ PERIPHERAL (*The investigated IUA is not a central aspect of the publication, but the publication contains evidence relevant to the review.*)
- ◯ NOT APPLICABLE (*The article reviewed provides no relevant evidence related to the IUA being investigated. Specifically, none of the numbered areas below are addressed.*)

1.X) IUA Notes

Use this space for notes and comments about how the publication reviewed addresses the IUA under investigation.

Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

2.0) Measurement and Methodology Issues

This section is primarily concerned with the scores produced by a test or derived through an analytical procedure. As these scores inform the IUA under investigation, evidence of the reliability/precision (or Intertemporal reliability) of the scores, and the overall methodological defensibility of the procedures used to generate those scores also provides support for the validity of the IUA (AERA, APA, & NCME, 2014; Kane, 2013).

For the purpose of this review, this section deals with the measurement and methodology issues of ***the process*** *used to obtain an estimate of teacher effectiveness from student standardized test scores*.

Note that evidence related to the measurement and methodology issues of the standardized tests that the students take is only marginally related to the IUA under investigation. The IUA under investigation must assume that the standardized tests given to students are sufficiently reliable and valid for its purposes. Thus, any evidence which seriously calls this assumption into question is worth noting, however evidence which supports the psychometric and methodological soundness of the student tests is insufficient to support the IUA.

This section contains three sub-sections:

- **2.1.0) Reliability/Precision and Stability**
- **2.2.0) Methodological Assumptions**
- **2.3.0) Construct Irrelevant Variance (CIV) / Bias**

2.1.0) Reliability/Precision and Intertemporal Stability

This sub-section deals with the scores produced to inform the IUA being validated. In the present study, the items in this sub-section relate to the reliability/precision or stability (also sometimes discussed as intertemporal reliability) of the teacher effectiveness estimates derived from student scores on standardized tests.

"The term reliability has been used in two ways in the measurement literature. . . . We use the term 'reliability/precision' to denote the more general notion of consistency of the scores across instances of the testing procedure, and the term 'reliability coefficient' to refer to the reliability coefficients of classical test theory" (AERA, APA, & NCME, 2014, p. 33).

RELIABILITY/PRECISION: "The degree to which test scores for a group of test takers are consistent over repeated applications of the testing procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group" (AERA, APA, & NCME, 2014, pp. 222-223).

INTERTEMPORAL STABILITY: "The extent to which scores on a test are essentially invariant over time, assessed by correlating the test scores of a group of individuals with scores on the same test or an equated test taken at a later time" (AERA, APA, & NCME, 2014, p. 223).

For the present review, this subsection deals specifically with the reliability/precision and intertemporal stability of the estimates of teacher effectiveness produced by the procedure used to derive them from student standardized test scores.

2.1.1) Addressed

To what extent was the reliability/precision, stability, or intertemporal reliability of *teacher effectiveness estimates* addressed by the publication reviewed?

- ○ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ○ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ○ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ○ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

2.1.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence **SUPPORTS** the IUA under investigation
- ○ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ○ (1) Evidence **CHALLENGES** the IUA under investigation
- ○ (NA) **Not Applicable** (*0 or 1 selected above*)

2.1.X) Notes

Use this space for notes and comments related to reliability/precision, stability, or intertemporal reliability. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

2.2.0) Methodological Assumptions

The items in this section deal specifically with the methodological assumptions of the procedures used to produce the scores on the "test".

In the case of the present review, these items relate to the assumptions of *the methods used to produce estimates of teacher effectiveness* from student scores on standardized tests.

2.2.1) Addressed

To what extent were the methodological assumptions of the procedures used to generate teacher effectiveness estimates addressed by the publication reviewed?

- ○ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ○ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ○ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ○ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

2.2.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence **SUPPORTS** the IUA under investigation
- ○ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ○ (1) Evidence **CHALLENGES** the IUA under investigation
- ○ (NA) **Not Applicable** (*0 or 1 selected above*)

2.1.X) Notes

Use this space for notes and comments related to methodological assumptions. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

2.3.0) Construct Irrelevant Variance (CIV) / Bias

Construct irrelevant variance (CIV) is a term used by Messick (1989) to describe factors that falsely inflate or deflate the measurement of a variable and therefore distort its interpretation, or distort its validity. As per Kane (2013) "[a]ny factor other than the construct that has an impact on the indicator (e.g., the method or context of observation) is a source of construct-irrelevant variance, or systematic error" (p. 40).

NOTE: Any discussion of CIV or bias in the scores, or in the procedures that produce the scores, should be noted in this section. If the publication being reviewed ALSO goes on to discuss possible ramifications of such bias, such as differential outcomes or possible adverse impact for members of affected subgroups, then that discussion should be captured in section 7.0 of this instrument.

2.3.1) Addressed

To what extent was CIV/bias addressed by the publication reviewed?

- ◯ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ◯ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ◯ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ◯ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

2.3.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ◯ (5) Evidence **SUPPORTS** the IUA under investigation
- ◯ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ◯ (1) Evidence **CHALLENGES** the IUA under investigation
- ◯ (NA) **Not Applicable** (*0 or 1 selected above*)

2.3.X) Notes

Use this space for notes and comments related to CIV/bias. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

3.0) Validity Evidence Based on Test Content

"Evidence based on test content that supports the intended interpretation of test scores for a given purpose. Such evidence may address issues such as the fidelity of test content to performance in the domainin question and the degree to which test content representatively samples a domain, such as a course curriculum or job" (AERA, APA, & NCME, 2014, p. 218).

"Important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA, APA, &NCME, 2014, p.14).

"Content-oriented evidence of validation is at the heart of the process in the educational arena known as 'alignment', which involves evaluating the correspondence between student learning standards and test content" (AERA, APA, & NCME, 2014, p. 15).

NOTE: After further discussion, and for the purposes of the present validation review, this area of validity evidence will focus mainly on the issue of alignment discussed in the standards. Evidence related to whether the construct of teaching is appropriately mapped by gains in student test scores will be captured in area 4.0, below.

This area of evidence must be considered in terms of the content of the analytical procedures used to estimate teacher effectiveness. As an allegory, imagine a test which only contains items requiring the addition and subtraction of fractions. Such a test could only support interpretations of the test taker's proficiency in the addition and subtraction of fractions, and it would be inappropriate to interpret its scores as a general measure of mathematics proficiency. Likewise, this section will contain evidence for or against the sufficiency of using student standardized test scores as the outputs of effective teaching. This section will ask questions such as, "Are student test scores a comprehensive measure of effective teaching?" and, "Are there any critical elements of teaching effectiveness which are not captured by student standardized test scores?"

NOTE: As a validation review, evidence may only be included in this section of the instrument if the author(s) of the paper being reviewed addresses the alignment of student standardized test scores to measure teacher effectiveness. Furthermore, this area of evidence will only be "directly studied" or "mentioned with evidence" if it relies on more than just logical argument, but presents empirical or external evidence to support the argument.

3.1) Addressed

To what extent is content-related evidence of validity addressed by the publication reviewed?

- ◯ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ◯ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ◯ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ◯ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

3.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence **SUPPORTS** the IUA under investigation
- ○ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ○ (1) Evidence **CHALLENGES** the IUA under investigation
- ○ (NA) **Not Applicable** (*0 or 1 selected above*)

3.X) Notes

Use this space for notes and comments about content-related evidence of validity. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

4.0) Validity Evidence Based on Response Processes

"Analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (AERA, APA, & NCME, 2014, p. 15).

"For tests that employ task presentations or response modes that are substantially different from those in the target domain (e.g., the uses of objective questions in cases where the performances of interest are extended), the possibility that irrelevant method variance will limit the validity of the proposed interpretation is a natural concern (Messick, 1989, 1994). If the processes involved in responding to the test tasks and to other tasks in the target domain are understood to some extent and if the test tasks seem to involve the same processes as most tasks in the target domain, extrapolation is likely to seem reasonable" (Kane, 2013, P. 28).

This area of evidence must be considered in terms of process of effective teaching. Are student standardized test score gains the desired outcome of effective teaching, or is it more complex or subtle than that? Are other response processes required to fully map the construct of effective teaching?

There is a great deal of potential overlap between this area of validity evidence and areas 3.0, above, and 5.1.0, below. Though subtle, this area deals with the *response process* of the "test" (i.e., do student standardized test score gains represent a sufficiently complex response with which to measure effective teaching?).

NOTE: After further discussion with the team about this validation review, evidence may only be included in this section of the instrument if the author(s) of the paper being reviewed addresses the adequacy or inadequacy of the response process of student standardized test score gains to measure teacher effectiveness. Furthermore, this area of evidence will only be "directly studied" or "mentioned with evidence" if it relies on more than just logical argument, but presents empirical or external evidence to support the argument.

4.1) Addressed

To what extent is response process related evidence of validity addressed by the publication reviewed?

- ○ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ○ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ○ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ○ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

4.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence **SUPPORTS** the IUA under investigation
- ○ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ○ (1) Evidence **CHALLENGES** the IUA under investigation
- ○ (NA) **Not Applicable** (*0 or 1 selected above*)

4.X) Notes

Use this space for notes and comments related to response processes. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

5.0) Validity Evidence Based on Internal Structure

Validity evidence in this section is primarily concerned with the theoretical construct that the test in question purports to measure, and the degree to which that measurement reflects the structure of the construct.

For the purposes of the present review, this section would refer to the theoretical construct of effective teaching.

This section contains two sub-sections:

5.1.0) Internal Structure
5.2.0) Reliability Coefficients / Internal Consistency

5.1.0) Internal Structure

"Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, NCME, 2014, p. 16).

Evidence based on internal structure concerns the degree to which the test or analytical process that supports the IUA reflects the theoretical construct on which it is based. This can include evidence of how faithfully the factor analyses, sub-scales, sub-tests, or various components of the testing procedure follow their hypothesized relationships.

NOTE: It is possible, though generally unlikely, that the present review will include evidence in this sub-section, as estimates of teacher effectiveness based on students' scores on standardized tests do not generally imply a specific or complex theoretical construct that might be tested with analyses of internal structure. If, however, the publication reviewed does discuss effective teaching as a theoretical construct and use factor analyses or other analyses meant to probe the relationship between components of the proposed construct, then the author(s) may introduce evidence relevant to this section.

5.1.1) Addressed

To what extent was the internal structure of the components (or scores) used to generate estimates of teacher effectiveness addressed by the publication reviewed?

- ◯ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ◯ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ◯ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ◯ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

5.1.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence **SUPPORTS** the IUA under investigation
- ○ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ○ (1) Evidence **CHALLENGES** the IUA under investigation
- ○ (NA) **Not Applicable** (*0 or 1 selected above*)

5.1.X) Notes

Use this space for notes and comments related to internal structure. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

5.2.0) Reliability Coefficients / Internal Consistency

"The term reliability has been used in two ways in the measurement literature. . . . We use the term 'reliability/precision' to denote the more general notion of consistency of the scores across instances of the testing procedure, and the term 'reliability coefficient' to refer to the reliability coefficients of classical test theory" (AERA, APA, & NCME, 2014, p. 33). Reliability coefficients and/or indicators of internal consistency for subscales or instruments that are used in the analytical procedures used to generate scores for the IUA under investigation are also included in this section. The most commonly reported reliability coefficient is Cronbach's Alpha, but split-half reliability, alternate form reliability, or generalizability coefficients may also be recorded in this section.

"In classical test theory, and in subsequent psychometric models, reliability is taken to be a necessary but not sufficient condition for validity" (Kane, 2013, p. 29).

NOTE: This section deals with the innards of a test, and as such, is very unlikely to come up in the present review. If the publication reviewed explores a complex theoretical construct related to effective teaching, any evidence provided is more likely to be recorded in 5.1.0.

5.2.1) Addressed

To what extent were reliability coefficients or indicators of internal consistency addressed by the publication reviewed?

- ◯ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ◯ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ◯ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ◯ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

5.2.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ◯ (5) Evidence **SUPPORTS** the IUA under investigation
- ◯ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ◯ (1) Evidence **CHALLENGES** the IUA under investigation
- ◯ (NA) **Not Applicable** (*0 or 1 selected above*)

5.2.X) Notes

Use this space for notes and comments related to reliability coefficients or internal consistency. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

6.0) Validity Evidence Based on Relations to Other Variables

"Relationships between test scores and other measures intended to measure the same or similar constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of different constructs provide discriminant evidence" (AERA, APA, & NCME, 2014, pp. 16-17).

"Evidence of the relation of test scores to relevant criterion may be expressed in many ways, but the fundamental question is always, how accurately do test scores predict criterion performance? . . .

Historically, two designs, often called predictive and concurrent, have been [used]" (AERA, APA, NCME, 2014, p. 17).

The literature has historically used several different terms for very similar sources of validity evidence. "The main idea is that measures of the same thing (construct) should correlate more highly than measures of [different] things" (T. Haladyna, personal communication, December 26, 2014). The terms that we use in this instrument for specific types of relationships to other variables are defined in 6.3, below.

6.1) Addressed

To what extent were relationships with other variables addressed by the publication reviewed?

- ◯ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ◯ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ◯ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ◯ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

6.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ◯ (5) Evidence **SUPPORTS** the IUA under investigation
- ◯ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ◯ (1) Evidence **CHALLENGES** the IUA under investigation
- ◯ (NA) **Not Applicable** (*0 or 1 selected above*)

6.3) Types

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, what types of evidence based on relationships to other variables did the publication reviewed provide?

- ☐ **CONCURRENT / CONVERGENT EVIDENCE** - The degree to which measures of the same or similar constructs correlate highly
- ☐ **DISCRIMINANT EVIDENCE** - The degree to which measures of different or unrelated constructs correlate poorly
- ☐ **PREDICTIVE EVIDENCE** - The degree to which a test or observation accurately predicts future performance or future outcomes (these future criteria may not be tests per se, but might be dichotomous outcomes such as college admission or four-year graduation, etc.)

6.X) Notes

Use this space for notes and comments related to relationships with other variables. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.0) Validity Evidence Based on Related Consequences

“The evaluation of test score uses requires an evaluation of the consequences of the proposed uses, and negative consequences can render a score use unacceptable… [although]… I am arguing for including evaluations of specific kinds of consequences (but not all consequences) under the heading of validity” (Kane, 2013, p. 46).

Kane (2013) identifies "three major categories of outcomes/consequences that have played and continue to play a major role in evaluating score-based decision rules: (1) the extent to which the intended outcomes are achieved, (2) differential impact on groups (particularly adverse impact on legally protected groups), and (3) positive and negative systemic effects" (p. 48). The consequences of a valid decision rule will achieve its intended goals at an acceptable cost, or will have an acceptable balance of intended and unintended consequences (AERA, APA, NCME, 2014; Kane, 2013).

In order to help the reviewers calculate the balance of intended and unintended consequences reported by the literature reviewed, the three categories provided by Kane (2013) will be split into four separate notes fields:

- **7.A) INTENDED OUTCOMES**
- **7.B) POSITIVE SYSTEMIC EFFECTS (unintended)**
- **7.C) NEGATIVE SYSTEMIC EFFECTS (unintended)**
- **7.D) DIFFERENTIAL EFFECTS / ADVERSE IMPACT (unintended)**

7.1) Addressed

To what extent were intended and/or unintended consequences of the IUA under investigation addressed by the publication reviewed?

- ◯ (3) **DIRECTLY STUDIED** (*evidence gathered and reported as part of the publication's study*)
- ◯ (2) **MENTIONED WITH EVIDENCE** (*e.g., cites other papers*)
- ◯ (1) **MENTIONED** (*no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence*)
- ◯ (0) **NOT ADDRESSED** (*mark "NA" below and skip to the next numbered area of evidence*)

7.2) Evidence

If DIRECTLY STUDIED ***or*** MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence **SUPPORTS** the IUA under investigation
- ○ (3) Evidence **NEITHER SUPPORTS NOR CHALLENGES** the IUA under investigation
- ○ (1) Evidence **CHALLENGES** the IUA under investigation
- ○ (NA) **Not Applicable** (*0 or 1 selected above*)

7.A) INTENDED OUTCOMES

Use this space for notes and comments related to the intended consequences of the IUA under investigation. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.B) POSITIVE SYSTEMIC EFFECTS

Use this space for notes and comments related to positive unintended consequences of the IUA under investigation that affect everyone. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.C) NEGATIVE SYSTEMIC EFFECTS

Use this space for notes and comments related to negative unintended consequences of the IUA under investigation that affect everyone. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.D) DIFFERENTIAL EFFECTS / ADVERSE IMPACT

Use this space for notes and comments related to unintended consequences of the IUA under investigation that affect certain sub-groups or populations differently than others. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

8.0) Other Validity Notes

"A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA, APA, & NCME, 2014, p. 21)

8.1) Overall Rating

According to the reviewer, and based upon the readings and discussions of the research team, taken in its entirety, how strong is the evidence of validity provided by this publication for the IUA being ivestigated in this review?

- ◯ (5) **SUPPORTIVE**
- ◯ (4) **SOMEWHAT SUPPORTIVE**
- ◯ (3) **NEITHER SUPPORTIVE NOR CHALLENGING**
- ◯ (2) **SOMEWHAT CHALLENGING**
- ◯ (1) **CHALLENGING**

8.A) Author's Concerns

What concerns or limitations do the author(s) raise about the IUA that have not already been captured in another part of this form?

8.B) Reviewer's Concerns

What concerns or limitations does the reviewer see that have not already been captured by another part of this form?

8.X) Other Notes

Provide other thoughts, comments, and notes here about the publication in its entirety or about issues not addressed elsewhere in this form.

0.X) Star Rating

On a five-star scale, how would you rate the overall value of this publication for inclusion on a list of recommended reading?

- ○ Five Stars
- ○ Four Stars
- ○ Three Stars
- ○ Two Stars
- ○ One Star

>>