

Running Head: CRITERION-RELATED EVIDENCE FOR ALTERNATE ASSESSMENTS

Criterion-Related Evidence for Alternate Assessments in Reading and Mathematics

Gerald Tindal  
Patricia Almond  
Josh Geller  
Paul Yovanoff

University of Oregon

10/30/06

Criterion-related Evidence for Alternate Assessments – 2

### *Criterion-related Evidence for Alternate Assessments in Reading and Mathematics*

The purpose of this chapter is to describe how information can be collected to provide criterion-related evidence supporting the claims of students' skill, ability, or trait while using specific measurement methods conducted in varying contexts. This type of evidence is particularly important for alternate assessments to acquire any value-added interpretations. Although content-related evidence may provide documentation about achievement on standards that is useful for peer review and accountability, it begs the question about the effect of performance (or proficiency) beyond the statewide testing outcome itself.

The first section provides an example of criterion-related evidence to illustrate the purpose for conducting the present study; the second section extends this example into an analytic framework; the third section presents preliminary and selective results from a study in middle-school mathematics; and finally, two important caveats in practice are considered.

#### *A Definitional Example of Criterion-Related Evidence*

Most skills, abilities, or traits are very complex constructs inferred from observing different behaviors in a number of different contexts. For example, the following methods could be used to measure reading skills (with some of them quite familiar in both large-scale tests and in classrooms):

1. Reading words from an unfamiliar text out loud for one minute with a parent volunteer
2. Orally recalling information from a familiar passage read independently silently.
3. Orally responding to teacher questions about a story read to the class.
4. Answering multiple-choice questions on a paper pencil bubble sheet in a state test.
5. Matching vocabulary words to definitions on a worksheet at the end of reading period.
6. Pointing to pictures of objects on flashcards when a teacher names them on a test.
7. Sequencing information from a passage with sentence cards in an alternate assessment.
8. Reading a paragraph and then independently summarizing it with a composition in class.
9. Predicting the ending to an unfinished' story in a small group discussion.

Notice that the first word in each sentence focuses on a different behavior, that the materials or methods rely on different measurement systems, and that these measures are completed in various contexts (conditions). The verbs range from selected responses (e.g., pointing, matching) to constructed responses (e.g., orally reading and recalling, explaining); the materials and methods generate different measures (e.g., words read per minute, answers to teacher questions, objects pointed to on flashcards); and the contexts (conditions) vary from independent work to small and large group participation. To what extent are these behaviors-methods-conditions related to each other and converge as measures of reading?

Criterion-related evidence addresses the degree to which different measures (behaviors-methods-conditions) relate to each other. Sometimes, these various measures are of the same or similar

10/30/06

Criterion-related Evidence for Alternate Assessments – 3

skills, abilities, or traits, and other times, they are of different skills, abilities, or traits. In the list above, all of behaviors relate to reading as a broad construct but differ in the specific aspects of reading that are being addressed. Certainly they use different methods and the conditions vary considerably. We hope that if reading is a coherent construct, then students would be consistently proficient across all nine measures (behaviors-methods-conditions). That is, the relation (or correlation) among the nine measures would be high. At the same time, we would NOT expect students who are proficient in these nine measures of reading to be skilled at playing football, running 100 meters, or even on another—but very different—academic skill such as calculating the answer to math addition facts or estimating outcomes in science experiments.

Which behaviors should be sampled, which methods should be used, and what contexts should be considered? And then, how should these behaviors-methods-contexts relate to each other and how should they relate to other known behaviors-methods-contexts of reading? To answer these two questions, measurement systems need to be devised that include (a) some type of prompt or stimulus, (b) a method for recording performance, and finally, (c) a metric to affix a (numeric) value on a scale. All three components then need to be operationalized in a context that allows for comparable interpretations across students, to specific skills, or over time to the learning of a student. Notice that these steps—the very operationalization of the behavior-method-context—are where many of the problems exist. Basically, the behavior-method-context defines the measurement system, which in turn needs to be considered in making interpretations of students' skills, traits, and abilities.

Measurement methods necessarily must be considered when making inferences from observations. No performance or behavior can be interpreted without considering the method used in documenting the behavior and valuing it with reference to either (a) a group of other students (norm-referenced), (b) a specific skill (criterion-referenced), or (c) prior performance over time (individual-referenced). For this reason, the *entire process* (behaviors-methods-contexts) used in measuring behavior needs to be carefully monitored before any analyses of outcomes can be conducted. Part of monitoring the process is to connect it to other similar and dissimilar *processes* (behaviors-methods-contexts).

Collecting criterion-related evidence increases understanding of the relations among various skills, abilities, and traits and the associated measurement methods (behaviors-methods-contexts). In collecting criterion-related evidence, multiple skills, abilities, and traits AND multiple measurement methods (contexts) need to be considered. It is not good enough to simply measure one skill, ability, or trait (e.g., predicting an outcome in an unfamiliar story) with two methods (e.g., using a multiple-choice response on a bubble sheet test completed independently and providing responses in a small group discussion in class). Nor is it sufficient to measure two different skills (e.g., identifying the main character and predicting an outcome in a story) with one method (e.g., using a multiple-choice response on a bubble sheet test completed independently). Rather, both are needed.

Criterion-related evidence helps determine both the integrity of the skills, abilities, and traits as well as the measures used to document performance; this kind of evidence becomes a critical component of the validation process. Using Kane's (1992) logic, validation is about making a claim (e.g., asking students to predict an outcome from a story reflects comprehension) and then

10/30/06

Criterion-related Evidence for Alternate Assessments – 4

supporting it with evidence, both procedural (describing the measurement methods and contexts) and empirical (statistically analyzing the outcomes). By documenting the behavior-method-context, procedural evidence is used to ensure the final outcomes and interpretations have reach beyond the immediate measures. By statistically documenting patterns and relations, empirical evidence is used to substantiate these interpretations.

*Measuring students with disabilities.* For students with the most significant disabilities, the behavior, method, and context need to be carefully considered. The behavior often is quite limited or also includes other interfering behaviors; students have a limited range of target behaviors (actual construct of interest) and often have competing and interfering access skills (irrelevant behaviors that are not of interest). The methods—reflecting how measures are developed, administered, and scored—must be flexible to accommodate the disability. And finally, the contexts must be realistic in reflecting transfer from school to home and community. To compound the problem, the empirical and statistical analyses often must be conducted with small sample sizes. At best, students with the most significant disabilities represent only 1% of the student population in schools, and they tend to be very different from each other in cognitive, motoric, and sensory proficiencies and needs.

With the appropriate procedures, we can then determine the relations between skills, abilities, or traits AND measurement of behaviors-methods-contexts to document performance and proficiency. Hopefully, we would find a similar pattern. For example, a student scores consistently by displaying various behaviors measured with different methods and in varying contexts. This kind of evidence would be convergent. It is important, however, to also consider divergent (or discriminant) information to provide a stronger validation argument, in which “irrelevant” information is found to be unrelated. Together, these two types of evidence support a validation argument necessary for a claim or inference. If the pattern of relations among the skills and methods is not as expected, then inferences must be reconsidered and sources of invalidity need to be determined. For instance, it is possible that unexpected patterns reflect construct irrelevant variance or construct misrepresentation and under-representation (see Haladyna & Downing, 2004).

#### *Multi-Trait, Multi-Method Analyses*

Perhaps the best model for understanding criterion-related evidence comes from Campbell and Fiske (1959) in their description of the multi-trait, multi-method analysis. [N. B. we translate the term “trait” to mean “skill” in this chapter]. In this process several different traits are measured using different methods to provide a correlation matrix that should reflect specific patterns (convergent or divergent or discriminant) that are supportive of the claim being made (that is, provide positive validation evidence). An important perspective in this process is that

. . . each test or task employed for measurement purposes is a “trait-method unit,” a unit of a particular trait content with measurement procedures not specific to the content. A fully rendered analysis is needed so that disconfirmation is possible as well as the obvious confirmation. If only one method is used, it is not possible to disentangle performance as a function of the method and trait. This two dimensional matrix provides reliability-related evidence (in each monomethod block) and validity evidence (in each heteromethod-heterotrait blocks). Ideally, the pattern of relations reflects a systematic ordering or relations such that the

10/30/06

Criterion-related Evidence for Alternate Assessments – 5

highest values are found with the same methods measuring the same traits, then different methods measuring the same trait, and finally different methods measuring different traits: “Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods” (p. 277).

In comparing various traits and methods, a correlational research design is used in which every student is tested in every skill and with every method:

1. Reliability—same method-same trait
2. Validity—different method-same trait (convergent within skills)
3. Validity—different method-different trait (divergent across skills)

The following argument is made on the basis of trait stability: Two different ways of measuring the same thing should be more consistent than the same way of measuring two different things. Notice that this argument can be made at different levels of specificity, whether the trait is considered reading versus math or decoding versus comprehension. This kind of analysis assumes that, for any given trait, most measurement methods are designed to be relatively robust and not significantly influence the outcome. Of course, this assumption is exactly what is being tested. If the construct of decoding depends upon how it is measured (the specific configuration of behavior-method-context), then the construct itself may be suspect and need to be redefined.

This analysis of relations with other variables begins to provide evidence for the argument (the claim) and prevents it from becoming a tautology. Basically, a “nomological net” is established in which various measures are constructed, each purporting to sample behavior from a domain and reflect a level of proficiency of skill or knowledge. When multiple measures converge in sensible ways, then inferences (of what is being measured as well as inferences of performance and proficiency levels) are supported. When the measures relate in ways that are not sensible, however, the claim needs to be adjusted and either the construct (trait) redefined, the measurement method redesigned, or the inference-interpretation reconsidered.

#### *An Illustrative Study of Criterion-Related Evidence*

The logic described above requires evidence to be collected at both levels of trait and measurement method. Also notice that it is important to actively provide negative (diverging or discriminant) validity evidence in this process. If two different measures (of the same trait) are used and one reflects proficiency and the other does not, it may be in (a) the type of behavior (trait), (b) the domain from which the measure is created (test specifications that define the trait), or (c) the scoring or judgment of proficiency (which is a function of the measurement method).

Following is a table describing three types of measures administered in a criterion validity study. It is important to note that this study is not a fully rendered version of the multi-trait – multi method matrix. Although we assessed students on three traits (language, reading, and mathematics) and two methods (rating scales and brief constructed performance tasks), traits and methods were not fully crossed. Language was only rated but no performance tasks were administered, and reading and mathematics were assessed only with performance tasks but not rated.

10/30/06

Criterion-related Evidence for Alternate Assessments – 6

In total, 62 students with the most significant disabilities (selected by teachers) were tested across several different states. All students were in either grades 7 or 8 and were tested in the spring of the year. See Table 1.

*Table 1. Measures Administered: Language Ratings, Reading and Mathematics Performances*

<i>Language Ratings</i>	<i>Reading Tasks</i>	<i>Mathematics Tasks</i>	<i>Mathematics Constructs</i>
Expressive- Labeling Objects	Signs & Symbols	1. Counting – Money	Money= Tasks 1+2+11+16
Expressive- Relaying Information	Letter Naming	2. Calculation – Word Story	
Expressive- Describing Action	Word Reading	3. Measurement – Compare Lengths	Problem Solving= Tasks 2+13+14+15+16
Expressive- Describing Events	Sentence Reading	4. Measurement – Identify Volume	
<u>Expressive Vocabulary</u>	Passage Reading	5. Counting – Number Recognition	Measurement= Tasks 3+4
Receptive- Attends to Others	Passage Comprehension	6. Counting – Number Recognition	Numbers= Tasks 5+6
Receptive- Follows 1 Step Command	Reading Total	7. Estimation – Time	
Receptive- Recognizes Objects		8. Estimation – Time	Time= Tasks 7+8+12
Receptive- Recognizes Action Needed		9. Tables and Graphs – Reading and Interpreting	
Receptive- Recognizes Attributes		10. Tables and Graphs – Reading and Interpreting	Tables and Graphs= Tasks 9+10
		11. Label Money	
		12. Label Time	Probabilities= Tasks 13+14
		13. Determine Probability Outcome	
		14. Selecting Division	
		15. Word Problem - Subtraction	Subtraction= Tasks 15+16
		16. Word Problem - Subtracting Money	
			Math Total

Because the full study exceeds the scope of this chapter, only selected findings are highlighted here to illustrate the logic of criterion-related evidence (using different behaviors measured in different ways and contexts and reflecting different metrics) and the logic of convergent and divergent evidence using a multi-trait – multi-method correlational analysis.

The main findings reflected an interesting relationship between the students' facility with language and their reading performance as measured across all operationalizations of reading. Notice that students using traditional language consistently perform better than those who are only beginning to use symbols and that these latter students consistently perform better than students who are pre-emerging or emerging. This relation became weaker, however, when reading did not require words but focused on ideas (comprehension). See Table 2.

10/30/06

Criterion-related Evidence for Alternate Assessments – 7

*Table 2. What is the Relationship between Expressive Labeling Objects Facility and Reading?*

Measure		Pre-Emerge	Emerging	Beginning	Traditional	Total
Signs & Symbols-N		1	2	8	38	
Average		4.0	1.0	6.1	13.6	11.7
Std Dev			1.4	4.5	3.9	5.4
Letter Naming-N		1	2	7	38	48
Average		2.0	2.0	15.1	19.3	17.6
Std Dev			2.8	6.8	1.5	5.2
Word Reading-N		1	2	6	38	47
Average		.0	.0	6.2	12.7	11.0
Std Dev			.0	5.1	6.8	7.3
Sentence Reading-N		1	1	7	38	
Average		4.0	.0	9.9	13.8	12.7
Std Dev		.	.	8.0	6.5	7.0
Passage Reading-N		1	1	6	38	
Average		.0	.0	23.0	27.4	25.6
Std Dev				12.2	12.7	13.5
Passage Comprehension-N		1	2	8	38	
Average		2.0	.5	3.0	4.4	4.0
Std Dev			.7	2.4	1.4	1.8

When the construct of language included a non-reading behavior, the differences between students with varying levels of symbolic usage were not much different, as one would expect. In other words, students who used traditional symbol systems attended to others equally well as those who used emerging or beginning symbol systems. See Table 3.

*Table 3. What is the Relationship between Attending to Others Facility and Reading?*

Measure		Pre-Emerge	Emerging	Beginning	Traditional	Total
Signs & Symbols-N		1	11	11	26	49
Average		4.0	8.9	10.3	13.8	11.7
Std Dev			5.4	5.9	4.4	5.4
Letter Naming-N		1	11	10	26	48
Average		2.0	16.4	16.2	19.3	17.6
Std Dev			6.7	6.6	1.4	5.2
Word Reading-N		1	10	10	26	47
Average		.0	11.5	11.2	11.2	11.0
Std Dev			7.7	8.2	6.9	7.3
Sentence Reading-N		1	11	9	26	47
Average		4.0	13.8	12.1	12.8	12.7
Std Dev		.	7.8	8.1	6.5	7.0

10/30/06

Criterion-related Evidence for Alternate Assessments – 8

Passage Reading-N	1	10	9	26	46
Average	.0	29.3	25.0	25.4	25.6
Std Dev		11.7	15.6	13.0	13.5
Passage Comprehension-N	1	11	11	26	49
Average	2.0	3.3	3.1	4.7	4.0
Std Dev		1.8	2.2	1.4	1.8

When the skill was mathematics and not reading, large differences appeared among the students with varying language communicative facilities. Students using traditional symbols to label objects performed much better in math overall than those who were only beginning to use symbols. This latter group also outperformed those who were pre-emerging or emerging. However, for some reason, attending to others also was related to mathematics at least between traditional and beginning (and unlike in reading where it was unrelated across the varying levels). The reason for this may be in the nature of the symbol system. While reading may have social connotations, mathematics does not; clearly, however, this conjecture needs more systematic research to ascertain the degree to which the claim can be supported. See Table 4.

*Table 4. What is the Relationship between **Language Communicative Facility** and **Mathematics**?*

Measure		Pre-Emerge	Emerging	Beginning	Traditional	Total
---------	--	------------	----------	-----------	-------------	-------

*Labeling Objects*

Math Total-N	1	3	8	38	50
Average	1.0	1.3	5.0	9.4	8.0
Std Dev		1.2	4.2	3.4	4.2

*Attending to Others*

Math Total-N	1	11	12	26	50
Average	1.0	6.6	6.1	9.8	8.0
Std Dev		4.1	4.1	3.6	4.2

Finally, in looking at a “trait” with high valence for both reading and mathematics, it appears that the size of a student’s vocabulary is generally related to their performance in these subject areas. Notice the sample size, however, when making this claim. See Table 5.

*Table 5. What is the Relationship between **Expressive Vocabulary** and (a) **Reading Comprehension** or (b) **Mathematics Facility**?*

Number of Words	0	1-3	4-8	9-15	16-25	26-50	50-200	>200
Comprehension-N	2	0	1	1	1	3	6	34
Average	1.5		5.0	.0	4.0	1.0	3.3	4.6
Std Dev	.7					1.7	1.0	1.5
Math Total-N	2	1	1	1	1	3	6	34
Average	1.5	2.0	5.0	.0	13.0	1.7	5.7	9.6
Std Dev	.7					2.9	2.5	3.3

10/30/06

Criterion-related Evidence for Alternate Assessments – 9

Finally, when looking at the relation between reading and mathematics, it seems that the two constructs are highly related. The correlation between various constructs within reading or mathematics are no more highly related with each other than they are across reading and mathematics. See Table 6.

*Table 6. What is the Relationship Within and Between Mathematics and Reading Performance?*

<i>Reading with Reading Total</i>		<i>Math with Math Total</i>		<i>Rdg Total with Math Total</i>
Signs & Symbols	.89	Money	.93	.89
Letter Naming	.91	Problem Solving	.88	
Word Reading	.93	Measurement	.79	
Sentence Reading	.97	Numbers	.81	
Passage Reading	.97	Time	.85	
Passage Comprehend	.85	Tables-Graphs	.79	
		Probabilities	.70	
		Subtraction	.77	

In summary, criterion-related evidence reflects the hypothesized relations among the same and different traits using the same and different methods. To the degree that the outcomes are consistent when they should be (measuring the same thing in the same ways), reliability evidence is being established. To the degree that different measures of the same thing are consistently related, convergent validity evidence is being established. Finally, to the degree that different ways of measuring different traits result in low relations, divergent validity evidence is being provided. This process usually proceeds by clearly articulating the construct being measured (including test content and specifications) and the method by which it is measured (providing procedural evidence), as well as using several different approaches to ensure the trait-method variance is not specific and can provide both convergent and divergent validity evidence.

#### *Caveats for Interpretation*

First, criterion variables need to include both those that might (should) be convergent in addition to those that might (should) be divergent. We included both in our illustrative findings. A common-sense perspective would dictate that in looking at the relationship between language and reading, one would expect a moderate (convergent) relationship to exist between labeling and (receptive) reading tasks (Table 2). Indeed, students with more traditional communication facility performed higher than students with beginning communication facility in all reading tasks, and in turn, this latter group performed better than students with emerging communication facility. Yet, we would not expect “attending to others” to be related to reading (rather, this variable should be the same across the four levels of communication facility). Indeed the findings in Table 3 indicate divergence from a linear trend, with the same levels of performance on the reading tasks reflected across three levels of communication (emerging, beginning, and traditional). Both of these basic findings were present in mathematics (Table 4).

Second, the type of measures can be either categorical or continuous. In our illustrative study, we used both: (a) the categorical variables included some language facility variables (labeling objects and attending to others) and (b) the continuous variables included both reading and

10/30/06

Criterion-related Evidence for Alternate Assessments – 10

mathematics tasks. Expressive vocabulary (Table 5) is somewhat in the middle with a continuous variable (number of words) organized into eight categories. Again, we found convergent results. Students with larger expressive vocabularies performed better in reading and mathematics.

Third, our measures had internal consistency that was very supportive of our constructs. Referring to Table 5, in reading, the subtests correlated highly with the total test; likewise, in mathematics, all subtests correlated highly with the total test. Unfortunately, the two total tests of reading and mathematics also correlated as highly with each other as the subtests within correlated with the total test. This is the heart of a multi-trait—multi-method matrix in providing both convergent and divergent evidence. Further research on the relations of the subtests with each other may help explain this finding. It is apparent that students with skill in reading also perform well in mathematics, even though the mathematics test was read to them and therefore should not have been an interfering and confounding access skill.

Two final caveats may be useful to consider when employing these results in a practical application. With positive criterion-related evidence, it is possible for the measures to be administered at different times, with one preceding the other even though the examples we presented focused on two measures given at the same time. Ideally, teachers would begin to develop formative measures that they could use well before the alternate assessment given at the end of the year. In this instance, we are considering the validity evidence as predictive and hope that the early performance information is consistently and highly related to the later performance information. Also, we focused only on measures of skills and not other variables that might be related to the development of those skills. We could have as easily considered other measures that should be related to the development of a skill or the display of a trait. For example, we could have measured the amount of instructional time and practice devoted to teaching mathematics skills with the expectation that it should be highly related to level of performance. We also could have measured the age of the student (as a proxy for instructional time) and would expect the same high relations with performance. With better information about contextual variables, it is possible to manipulate one and affect change in the other. Although this logic is correlational and should not be confused with causation, it does begin to lead to stronger research designs as well as bring focus on variables that can be manipulated by teachers in the classroom (e.g., spend more time to see growth).

10/30/06

Criterion-related Evidence for Alternate Assessments – 11

### *References*

- American Educational Research Association, American Psychological Association, and National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait, multi-method matrix. In W. A. Mehrens and R. L. Ebel (Eds.), *Principles of educational and psychological measurement: A book of selected readings* (pp. 273-302). Chicago: Rand McNally.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527-535.