Generalizability Theory Applied to Reading Assessments for Students with Significant Cognitive Disabilities

Gerald Tindal

Paul Yovanoff

Josh Geller

University of Oregon

Abstract

Students with significant disabilities must participate in large-scale assessments, often using an alternate assessment judged against alternate achievement standards. The development and administration of this type of assessment must necessarily balance meaningful participation with accurate measurement. In this study, generalizability theory is used to estimate the dependability of reading items and tasks that have been administered using two formats of receptive and expressive communication. The results reflect a trade off between meaningful participation and accurate measurement of students with significant cognitive disabilities, particularly when considering these two formats. Significant variance occurs for persons interacting with tasks while the effect of raters is negligible. Furthermore, these results appear to vary across administrative format.

Generalizability Theory Applied to Reading Assessments for Students with Significant Cognitive Disabilities

States are faced with unprecedented pressure to include students with the most significant disabilities in large-scale testing programs and have their scores count in making Adequate Yearly Progress (AYP). AYP Measurement requirements include, (a) developing meaningful assessments aligned with state standards, (b) administration that provides fair access for a population with various behaviors that often interfere with valid traditional testing (e.g., require assistive technologies, scaffolds, diverse communication needs), and (c) scoring performance in a manner that accurately scales proficiency levels. These minimal requirements render the development and use of appropriate tests extraordinarily challenging.

In this process, the emphasis needs to necessarily balance meaningful participation and generation of consistent and accurate outcomes. These two requirements, however, often are in contradiction. To make participation meaningful, the administration of assessments requires flexibility and thereby may compromise the standardization needed to make comparable judgments of proficiency. Standardized instrumentation and administration by definition is less flexible and may result in less meaningful participation for students with significant disabilities if the standardization ignores varying access skills. Without careful measurement development, these considerations may present a dilemma with respect to conventional standards for psychometric validity and reliability. More valid administration may attenuate measurement reliability.

Development of the alternate assessment is structured generally in terms of sampled observations from measurement conditions, e.g. populations, items, domains, formats, etc, Although we may *fix* one or more measurement condition(s) to increase precision, this kind of restriction results in standardizaiton and also limits the measurement condition(s) to which generalizations can then be made. As Brennan (2001) writes “in other words, fixing a condition of measurement reduces error and increases the precision of measurements, but it does so at the expense of narrowing interpretations of measurements” (p. 2). This language “provides an elegant explanation of the reliability-validity paradox, whereby attempts to increase reliability through standardization (i.e., fixing facets) can actually lead to a decrease in some measures of validity (Lord & Novick, 1968, p. 334 as cited by Brennan, 2001, p. 132).

Portfolios and performance assessments. The potential tradeoff between reliability and validity is central to the development and administration of any assessment for students with significant disabilities. In order to measure their performance, the instrument often must become less standardized, thereby threatening reliability. Probably the easiest and most direct reflection of this paradox is the use of *portfolios* (collections of evidence and work samples) versus the administration of *performance assessments*. These two options are probably the most widely adopted methodologies in alternate assessments.

Portfolios are highly flexible, and have the advantage of allowing teachers to ‘customize’ the kind of tasks being used to demonstrate proficiency and, in the process, rely on behaviors within the student’s repertoire. In a sense, accommodations are built into the administration process. For example, mathematics skills may be assessed in a number of different ways that use manipulative objects, paper-pencil problem sheets, and interactive tasks in which the student responds with scaffolds. The assessment provides a product from any of these behaviors that can then be accompanied with either explanatory or interpretive comments. Although this collection of evidence is flexible, it also provides problems with consistency of interpretation. When the work samples are different and the teachers collect them in different ways, it is almost impossible to come to common judgments. Furthermore, these work samples are collected over time (during the course of a year), making it difficult to compare them with each other.

In contrast, performance assessments tasks are more standardized in which items have been developed a priori following some kind of task specifications, the teacher follows a general protocol governing how to present the problem, directions are used to prompt the student, and performance is scored as the student interacts with the prompt. Although it is easier to compare one student to another, it is difficult to ensure meaningful participation. Many students with the most significant disabilities require some kind of assistance or accommodation, and to the degree that this is not offered (or it is offered differently to different students) it is impossible to determine whether or not performance is impeded. Even if accommodations are allowed, such changes (in the way the test is given or taken) must not change the construct being measured and thereby jeopardize the measurement validity.

In summary, these two methodologies present a difficult problem of creating alternate assessments that reflect meaningful participation, and systematic performance variability. Ideally, observed scores should result from true differences among students. Observed performance variability, however, may arise from at least three additional systematic sources of measurement error: (a) tasks, (b) occasions, and (c) the scorers. Tasks, occasions, and scorers are the sources of variability external to true student proficiency. If variation in the kinds of *tasks* being used in either a portfolio or performance assessment effect observed student variability, it would be important to know this so that adjustments could be made (for example, develop more clear test specifications or better task development). If observed scores vary in relation to the schedule of administration over *occasions*, it would be critical to make the earlier and later assessments more comparable. Finally, if teachers vary greatly in their scoring of performance, it would be important to make them more consistent (in harshness or lenience).

These three common ways in which construct-irrelevant variation (error) enters an assessment system (tasks or items, occasions of events, and raters or scorers) can be studied through generalizability theory. In the remainder of this paper, we provide a brief explanation of generalizability theory, then we review some commonly found outcomes pertaining to performance assessments, and finally, we describe the highlights of a study in which generalizability theory was used to understand the variation from tasks, format, and raters for a reading test administered to students with the most significant disabilities.

Generalizability Theory

Generalizability theory (G theory) extends classical test theory by isolating the systematic variation due to students, items, occasions, and raters (or any other variable that appears to create variation in observed scores). Classically, student observed scores are partitioned into two parts, (a) true scores, and (b) error scores. While student true score variability is desirable, all other variability is considered irrelevant, and therefore regarded as 'error' variance (attributable to sources such as items, occasions, raters). G theory focuses closely on error variance, and extending beyond classical test theory, provides estimates of how much effect items, occasions or raters have on the observed score variability. Ultimately, using these estimates of error variance, G theory provides various indices of measurement reliability.

According to G theory, sources of error are called '*facets*' of the measurement condition. The theory conceptualizes the student observations as samples or replications from a '*universe*' of similar measurement conditions composed of all possible combinations of the facets. Replicated observations may vary randomly or systematically. Is the variability due to random student differences, or systematic processes, perhaps attributable to irrelevant sources of error, e.g. the measurement facets? Using analysis of variance as a statistical model, G theory estimates error variance associated with the measurement facets. "What is considered unexplained error in classical test score theory may be portioned into distinct components in G theory" (Brennan, 2001, p. 54).

The universe of admissible observations. Although a number of different facets may be of interest, three have dominated the literature, (a) tasks or items (denoted as *i* or *t*), (b) occasions (denoted as *o*) and (c) raters (denoted as *r*). These facets are considered conditions of measurement that are defined by the researcher as the '*universe of admissible observations*'. In other words, replications of the student performance are sampled from a clearly specified measure process that is constrained by the admissible observations. If the facets are fully crossed, then the universe of admissible observations includes all possible combinations of the facets. The facets may be nested, in which case only some combinations constitute the sampling space, restricting generalizability inference. Note, though students are explicitly modeled as a source of variance they are not a facet. They are regarded conventionally as the '*object*' of measurement, sampled from the population of students. Considering all possible combinations, students and the measurement facets are used to explain the observed score variability in terms of true score variance and error variance (differentiated with respect to the facets).

The universe of generalizability. The ability to examine the reliability of decisions that are made from observed student performance is the advantage to differentiating and estimating the sources of systematic error variance. Variance estimates obtained in a G study can be used to construct efficient measurement and decision-making systems. In a Decision study (D study), the investigator considers various sampling configurations with respect to the measurement facets, e.g., numbers of items, occasions, and raters. Each configuration defines a *universe of generalizability*. For each universe of generalizability, the D study results include estimates of universe scores (true scores), and reliability of observed scores. The estimated reliabilities are useful for tailoring efficient and acceptably accurate measurements. Two types of reliability estimates are provided, (a) the generalizability coefficient for relative decisions, and (b) the coefficient of dependability for absolute decisions. Relative decisions pertain to inferences about performance relative to other students, while absolute decisions relate to mastery-type classifications.

G studies of performance assessment. Performance assessments are very well suited to the use of generalizability theory in understanding the influence of tasks, raters, and occasions on estimates of performance (Brennan, 1996, 2000). A study by Shavelson, Baxter, & Gao (1993) provides a prototypical example of this kind of research in which students completed five independent tasks and responded to a series of questions that were evaluated by a team of teachers using a rating scale to evaluate quality. As summarized by Brennan (2001), "the G-study estimated variance component for persons $\sigma^2(p) = .298$ is relatively large, but the estimate variance component for the *pt* interactions $\sigma^2(pt) = .493$ is even larger. By contrast, $\sigma^2(r)$, $\sigma^2(pr)$, and $\sigma^2(rt)$ are all close to zero, which suggest that the rater facet does not contribute much to variability in observed scores" (p. 118). In fact, this finding is quite typical.

1. Shavelson, Baxter, and Gao (1993) reported "the major source of measurement error was due to person x task x occasion interactions (59% of the variability)… the second largest source of error variance was the person x task interaction (32% of the total variability)" (pp. 223-224).

2. Lane, Liu, Ankenmann, & Stone (1996) report that “the variance component for the person x task interaction accounts for the largest percentage of the total variance” (p. 80).

3. Gierl (1998) reported that “the largest variance component was attributed to persons” (p. 95).

4. Gao and Brennan (2001), when studying the sampling variability of variance components across studies, reported that tasks were consistently the most notable source of error variance, and that error associated with raters was minimal. Also, the reliability of performance tasks appeared to depend on content area (listening or writing). Perhaps most noteworthy, the investigators report stable variance and reliability estimates across various studies.

5. Hintze and Pettite (2001) found that the greatest amount of total variance explained (62%) was due to “individual variation or differences among the participants” (p. 164) with only 15% of the variance due to setting, 6% due to repeated measurement over time, and 1% due to the interaction of setting with occasion.

6. Bruckner, Yoder, and McWilliam (2006) studied the measurement of preschoolers with grammatical and phonological impairments using G theory. They found that student scores across raters were reliable, but the session-by-student effects indicated unreliability associated with session (occasion). In their research, the session or occasion was actually the replicated task. Their results could be alternately interpreted as a task-by-student effect, indicating that task sampling is a possible problem.

In summary, researchers are consistently finding that variation in performance arises from the persons (students) who are participating and that this variation further interacts with tasks and occasions. Rather than placing all non-student variation into an overall error term, it can actually be partitioned into any of these facets. The findings suggest that student relative standing varies from task-to-task, and to a lesser extent across occasions. One consistent finding is the raters tend to be a small source of error, and this may be explained by the quality of rubrics and training upon which ratings are structured.

In the next section of this paper, we report on a recent study that illustrates application of generalizability theory to the investigation of assessment of reading performance for students with the most significant disabilities.

Methods

In generalizability research, the design of the study is critically important as it determines the interpretations or generalizations that can be made. In this particular study, some facets were deemed fixed while others were random. "Although the power of generalizability theory is most likely to be realized when a G study employs a fully crossed design (Brennan, 2001, p. 17), we were interested in two formats (receptive and expressive communication) for each student and therefore, used a mixed model. Likewise, the procedures determine the confidence with which the results can be trusted; therefore, we devote most of the discussion to these two topics and then discuss the measures, their administration, scoring, and rating on a construct of reading.

Design

Conducting generalizability research requires a careful data collection design enabling the estimation of relevant variance components. Similar to experimental research with accurately specified variables, we used a data collection system that controlled for the measurement facets by sampling factors listed in Table 1. Consistent with the G study nomenclature, we conceptualize the *universe of admissible observations* to include (a) any child eligible for participating in the alternate reading assessment, (b) responding to any items randomly sampled from a fixed set of reading tasks, (c) comprising form A or B, (d) of expressive and receptive administration formats, and (e) scored by any of a sample of raters. Additional sampling variables not explicitly included in the G study are listed in Table 1, e.g. occasion, teacher, state, Some sampling factors are not included in the G study because of small samples or lack of direct relevance to the research questions.

The G study focuses specifically on person (*p*), item (*i*), and rater (*r*) variance components. The universe of admissible observations includes any student taking any item nested in specific tasks administered in each of two specific formats. These performances are then scored by any trained rater. Because of the nested data structure a series of G studies are completed independently for the two administrative formats (*a*) by the six reading tasks (*t*). The G study design is summarized as: $(p \times i{:}(a \times t) \times r)$.

The G study design does not take into consideration all aspects of our data collection and measurement development. Some analyses are not possible with the sample sizes and the confounding nature of the sampling, e.g. items are nested in administrative formats by task, and students were randomly assigned to one of two forms A or B. Independent G and D studies were completed for each combination of two administrative formats and six tasks. Also, a separate study focused on form equivalence across the 12 combinations.

Table 1. Generalizability Study and Sampling Design Specifications

Measurement Facets and Sampling Factors	N	Random/Fixed	Sampling Notes
		G Study Facet	
Persons (*p*)	n_p=81	*Random* sample from student population	The population is the *object of measurement*, and is not considered a design facet,
Task (*t*)	n_t =6	*Fixed* restricted to specific reading skill domains	Fully crossed with persons
Item (*i*)	n_t = varies	*Random* sampling from infinite universe of items	Sampling is the conceptualization of item preparation; n_i varies per task
Administration Format (*a*)	n_t =2	*Fixed* restricted to Expressive and Receptive	Expressive/Receptive Fully crossed with persons and tasks
Form (*f*)	n_f =2	*Random* sample A/B from an infinite universe of forms	Multiple ‘parallel’ forms were created by expert staff item development
Rater (*r*)	n_r=5	*Random* sample 1, 2, . . ., 5 from an infinite universe of raters	All forms rated by each of 5 raters; Sampling is the conceptualization for rater training, rubric preparation, etc.
		Additional Sampling Factors	
Occasion	n =2	*Random* sample 1, 2 of an infinite set of occassions	
Teacher	n=66	*Random* sample 1, 2, . . ., 66 of an infinite set of teachers	Sampled within states
State	n=7	*Random* sample 1, 2, . . ., 7 of a conceptually infinite set of states	Volunteer recruitment through personal contacts

Procedures

Our procedures included sampling teachers from various states who identified students for research participation. Teachers were asked to administer two measurements (expressive and receptive) within one week. Using random assignment, teachers were asked to administer either of Form A or form B of the measures. As described below, measurement development, administration, scoring, and rating were conducted in an effort to obtain sufficient data for estimation of item, task, and administration format variance components. Using these variance estimates, reliability estimates for a viable measurement designs (sampling configurations) were computed in Decision studies.

Students. Participants in the generalizability study included students sampled from seven states: Alaska (1 teacher with 1 student), Iowa (9 teachers with 15 students), New Mexico (7 teachers with 10 students), Oregon (4 teachers with 8 students), Utah (13 teachers with 22 students), Washington (6 teachers with 9 students), and West Virginia (14 teachers with 16 students). A total of 54 teachers and 81 students took part in this study. Students were allowed to self-identify ethnicity; responses were consolidated into six approximate categories. The students were predominantly Caucasian (66%). Fourteen students were Hispanic (17%), three students were African American (4%), three students were Native American (4%), two students were Asian (3%), and the remainder was unidentified. Males made up 67% of the sample, females 33%. Each student in the sample was a student with significant disability. Disabilities represented in the sample were predominantly autism (~12%), mental retardation (~22%), and other health impairment (~35%).

Measures. Six reading skills were measured. Figure 1, illustrates the Form, Task, and Item structure of our data collection. All forms within format have the same number of items.

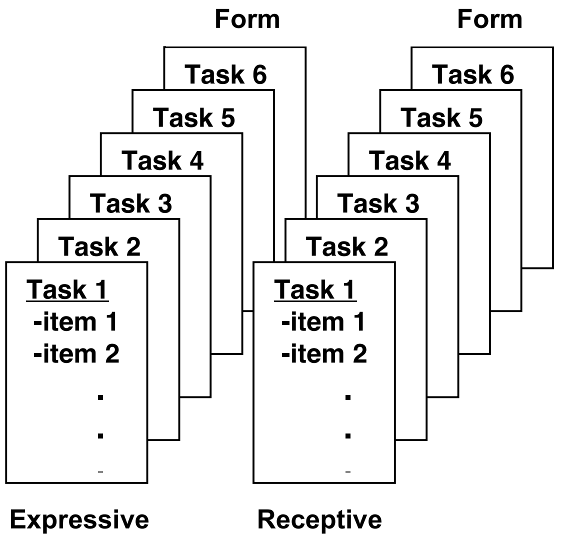

Figure 1. Form, Task, and Item Sampling Structure

Six tasks, each with approximately 10 items, were constructed. Each task was constructed with two forms, and also in an expressive and receptive administration format. Table 2 provides a list of skill domains and numbers of items per measurement. Forms A and B were identical with respect to numbers of items.

Table 2. Number of Items per Task by Format

	Administration Format	
Task	Expressive	Receptive
1. Identify Signs and Symbols	10	10
2. Letter Naming	10	10
3. Word Reading	10	10
4. Sentence Reading	5	5
5. Passage Reading	6	6
6. Comprehension Passage	6	6

Administrative format. We began the study by considering two different formats for administering alternate assessments: expressive and receptive. We chose these two formats because of the need to vary assessments for students who may or may not use traditional communication systems (marking test booklets or speaking).

Indeed, a sizable group of students with the most significant disabilities may need communication boards, use Braille, require pointing responses with a joy stick, be able to make an eye blink only, or require any number of different assistive technologies that 'activate' a response that reflects an answer to a prompt or task. We have labeled this mode of communication as *receptive*.

In contrast, many students with significant disabilities communicate in the traditional manner, speaking and completing production tasks (writing words or solving math problems). They can communicate actively in structuring a response. We have labeled this mode of communication as *expressive*.

We have labeled this dichotomy of receptive and expressive communication as 'administration format' (*a*), and include it as a measurement facet to determine if the student performance is generalizable across format. Will the variation from an assessment be the same from these two administration formats? We also were interested in variation from using different reading tasks (*t*). Will the variation from administering various reading tasks be similar or different depending on the type of task being presented. For this facet, we considered the following five reading tasks: (1) identifying symbols, (2) naming letters, (3) reading words, (4) reading sentences, (5) reading passages and answering comprehension questions. Because we had to present each student with two administration formats (receptive and expressive) we had to develop two forms (*f*) (of different items) so that we did not repeat the exact item twice. Therefore, form is nested in format (*f*:*a*): Each student received one form (e.g., Form A) using a receptive format and another form (e.g., Form B) using an expressive format. To control for form and administration format, as well as order of administration, we randomly assigned students to these conditions and then randomly assigned the administration order. After the administration, we trained five raters (*r*) to evaluate performance on a holistic rubric so we could ascertain the rater generalizability.

Teacher administration, measurement form, and scoring. For administration, teachers were randomly assigned to (1) order of administration of expressive/receptive formats, and (2) form A or B of the two formats. Table 3 indicates the numbers of teachers in the various assignments. All materials were sent to teachers with scripted instructions for administration; the research team scored all materials once teachers completed the testing and sent in the materials.

Table 3. Number of Participants Randomly Assigned to and Counter-Balanced for Combinations of Communication Format (Receptive/Expressive) by Forms (A/B)

Format (r /e) by Form (A/ B)	N	%
rA/eA	20	24.7
rA/eB	19	23.5
rB/eA	26	32.1
rB/eB	16	19.8
Total	81	100.0

All students completed each task 1 through 6 in both of the two formats, (expressive, receptive). Students were administered both forms sequential with approximately one week between administrations. Teachers scored each item per task upon administration and immediately returned the scored performance for analysis.

Rater scoring. While teachers scored each item within each of the six tasks, raters were asked to consider the entire set of six performances. Receptive and expressive assessments were each scored by five independent raters. All raters were pursuing Master's degrees at a research institution. All raters were trained by the researcher in the scoring procedures. Four participated in two weekly meetings to ensure consistency while the fifth rater scored all assessments independently. All expressive and receptive assessments were randomized and numbered (1-n). Expressive and receptive assessments were separated for scoring. Assessments were scored with like assessments (e.g., all Expressive protocols were scored together). Within the assessment type, the forms were randomly distributed in the scoring to avoid scoring bias. Table 4 provides the rating scale used when scoring each student.

Table 4. Rating Scale for Rating Overall Student Performance Across All Six Tasks

0	1	2	3	4	5
Student demonstrates no behavior.	Student demonstrates virtually no comprehension and few symbol, letter, or word skills.	Student demonstrates limited comprehension with some symbol, letter, and word reading skills.	Student demonstrates emerging comprehension with some sentence reading and few passage reading skills.	Student demonstrates basic comprehension with accurate word and sentence reading skills, and some passage reading skills.	Student demonstrates full comprehension , including accurate symbol, letter, word, sentence, and passage reading.

To practice rating, each rater received the same expressive and receptive assessments. Each rater independently reviewed and rated the assessments using the rating scale. Raters shared their score with the group and gave their justification for how they rated a particular assessment task. Discussion ensued and a group agreement was reached as to the appropriate score to assign the student's assessment. This process continued with the other forms until all four assessment types had been discussed (Expressive A, Expressive B, Receptive A, and Receptive B).

Once scoring agreement was set, raters were directed to score the remaining protocols without conferring with each other. The rules for evaluating were proposed as follows. The raters reviewed his or her entire stack of assessments with no judgment. Next, they holistically rated the tasks to get a general overview of the student's skills and knowledge. Raters then considered patterns of behavior within and across tasks that made up the assessment. Finally, the rating scale was applied and student performance on the assessment was assigned a number from the rating scale, 0-5. They spent no more than 5-7 minutes evaluating each assessment.

After one week, the researcher met with all raters to discuss any questions raised or inconsistencies noted during the rating process. At this point, raters received another set of assessment tasks. The rating process was repeated for the second week until all four raters had scored every assessment. The fifth rater scored all assessment tasks independent of the group interaction.

Results

The data were initially analyzed at the item and task level with the results reflecting a number of interesting findings.

1. The receptive format was generally easier for this population of students. Very few receptive tasks averaged below 50% correct while very few expressive tasks averaged above 50%.

2. The different tasks were not always uniformly easy within either of the two formats (mode of communication): Naming letters was quite easy while reading words and passages, as well as answering comprehension questions were all quite difficult.

3. Some item tasks interacted with format: Two tasks on Form A Expressive were about the same level as the other item tasks for receptive (marked in bold in Table 5).

4. Finally, according to the descriptive statistics in Table 5, the standard deviation for some tasks was quite high, indicating that group of student performance on this task is relatively variable. Indeed, this same finding can be seen in the levels of reliability obtained using classical analyses. All tasks except the two comprehension tasks for the receptive format were above .85 using inter-item correlations (and many were above .90). This outcome may well be a function of the bi-modal distribution of the group with some clearly low and some clearly high in their

performances (and few in the middle); this configuration inflates coefficient alpha in which items are inter-correlated with other items and the total test score.

Table 5. Task Proportion Correct Descriptive Statistics by Format by Form

	Expressive				Receptive			
	A		B		A		B	
Task	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Sounds and Symbols	0.47	0.27	0.28	0.25	0.71	0.35	0.64	0.33
Letter Names	**0.80**	0.28	0.57	0.44	0.77	0.37	0.76	0.35
Word Reading	0.33	0.34	0.22	0.34	0.47	0.39	0.53	0.38
Sentence Reading	0.45	0.39	0.31	0.38	0.60	0.41	0.56	0.40
Passage Reading	**0.43**	0.39	0.33	0.40	0.41	0.37	0.46	0.34
Comprehension	0.17	0.27	0.14	0.27	0.55	0.33	0.39	0.31

Because of this outcome, a G study is even more informative. It is designed to help understand the source of error variance associated with tasks. Specifically of interest is the extent to which the reliability of observed student performance is diminished by (a) items within tasks, (b) items within administration format, and (c) rater within administration format. Based on the error variance estimates from the G study, a series of D studies, were designed to estimate the measurement reliability for various measurement designs (hypothetical combinations of tasks, formats, and raters). Twelve independent studies were completed for each of the six tasks by the two administration formats.

For each D study, the generalizability coefficient was estimated for measurement designs with between 3 and 10 items. The *generalizability coefficient* (G coefficient) provides an index of reliability consistent with classical measurement theory. Specifically, the G coefficient is an estimate of the proportion of observed score variability that is attributable to the true score variability. Higher coefficient values for specified configurations, i.e., number of items and number of raters, indicate higher measurement reliability in terms of relative student order. For instance, does student ranking depend on the item, format, or rater facets? Findings for this type of question are consistent with norm-referenced interpretations of performance.

These coefficients have been reported in Table 6 and indicate that the amount of variance accounted by persons alone or persons interacting with items, which was typically greater than the variance accounted by items (and was quite consistently negligible). Furthermore, when comparing expressive versus receptive formats, the variance accounted by persons interacting with tasks was greater with receptive tasks than expressive tasks, except for naming letters. This means that the format of the item task had a significant influence on the variance for *p* x *i* (persons interacting with items). This finding has an important implication in that, though receptive tasks may result in higher average performance, this performance has more variance that cannot be explained until specific reference is made to the person and the item. Typically, this variance accounted for between 11% and 16% of the performance.

Table 6. Item Expressive and Receptive Format Random Effects G Study Variance Component Estimates

Effect	E-SS	R-SS	E-LN	R-LN	E-WR	R-WR	E-SR	R-SR	E-PR	R-PR	E-PC	R-PC
person (p)	.049	.094	***.149***	***.120***	.097	***.131***	.125	***.142***	.087	.102	.036	.081
item (i)	.012	.004	.001	.000	.001	.004	.000	.001	.002	.001	.003	.008
p x i	.015	***.127***	.066	.059	.068	***.117***	.065	***.103***	.060	***.145***	.066	***.162***

E=Expressive and R=Receptive
SS=Sounds and symbols, LN=Letter names, WR=Word reading, PR=Passage reading, PC=Passage comprehension

Knowing the student and facet component estimates from a G study, we are able to use the results from a D study to help us develop assessments with sufficient items so that we can achieve a reliable outcome. In the two charts below, we report the generalizability coefficients for various numbers of items in each of the six tasks.

We began with as few as three items and included as many as 10 items (some tasks though never actually had this many items). As can be seen in Figure 2 and Figure 3, the generalizability coefficients were lower for receptive tasks than for expressive tasks, for any given number of items, with the exception of signs/symbols. Furthermore, for all tasks, the generalizability coefficients were surprisingly high even with as few as 3-5 items. Increasing the number of items beyond 5 rarely resulted in (practically) significant improvements in the generalizability coefficient. Finally, when considering the number of raters needed to achieve high generalizability coefficients, either 2 or 3 would be sufficient, beyond which only small increases would be found. For expressive tasks, these coefficients were typically higher for decoding tasks (naming letters, reading words, sentences, and passages); however, when the task focused on meaning (signs/symbols and comprehension), more items would be needed. The same was true for receptive format, though all tasks required more items to achieve sufficient generalizability coefficients.

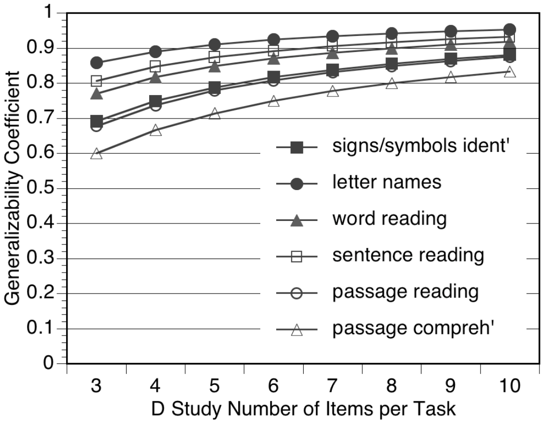

Figure2. Receptive Reading Task Format D Study Generalizability Coefficient Estimate

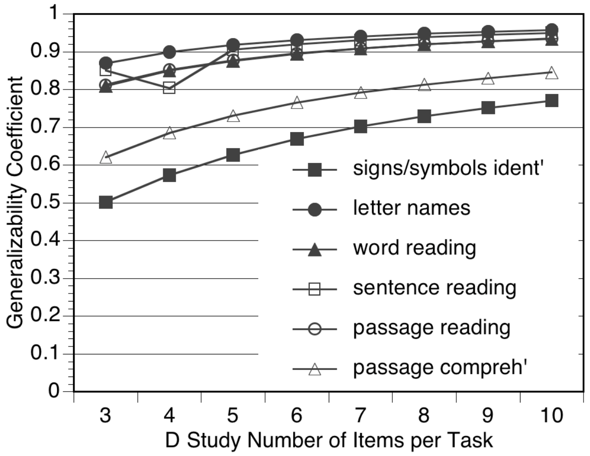

Figure 3. Expressive Reading Task Administration Format D Study Generalizability Coefficient Estimates

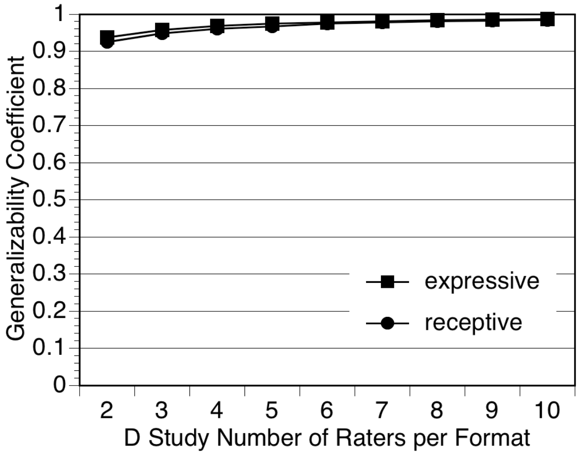

Figure 4. Rater D Study Generalizability Coefficient Estimates

Discussion

Preliminary item analyses indicate that the measures functioned well. Like any test, performance is likely to vary; the measures showed that students were indeed learning to read with a variety of skills being reflected in their performance on the variety of tasks. Both expressive and receptive formats showed these varying levels of skill in reading. Forms A and B were equivalent in difficulty for Expressive format, but not for the Receptive format.

An expected finding was that performance was higher on the receptive tasks. We developed the tasks in order to measure a target skill, namely reading, with the prediction that many students with the most significant disabilities need accommodations because of behavior/skills that interfere with traditional testing. Many students in this population, for example, have limited capacity to express themselves and must rely on nodding, blinking, eye movement, or other behaviors to reflect their choice. By administering a reading test that allows this response, we believed it possible to measure a level of performance unimpeded by limited access skills. Indeed, such was the case. If we define reading even from the viewpoint of the National Reading Panel (requiring that decoding be considered part of the construct), it is possible to measure it with a receptive format (e.g., when I say a word, please point to the card that displays that word).

An unexpected finding was that estimated variance tended to be smaller for expressive tasks, particularly in the interaction between persons and items. The generalizability results indicated that the expressive format items are usually more reliably sampled, though this depended somewhat on the reading task (domain) being sampled. For example, in ‘word reading’, ‘passage reading’ and ‘passage comprehension’, the expressive format resulted in higher reliability than the receptive format. ‘Letter names‘ and ‘sentence reading’ were both reliably sampled, with approximately 3 items being adequate for a generalizability coefficient of .85. In contrast, for ‘signs and symbols identification’ were poorly sampled in both formats though expressive item sampling was slightly less reliable.

Like much of the previous research on generalizability, rater reliability was extremely high for both the expressive and receptive formats. As Linn and Burton (1994) wrote over a decade ago, “quite high levels of generalizability across raters can be achieved when well-defined scoring rubrics are reinforced by intensive training and ongoing monitoring during rating sessions” (p. 5). In our study, two raters resulted in generalizability coefficients of approximately .90. Basically, the variance was primarily in the item task and in the interaction of the item task with persons, not in the ratings of their performance. In part, this finding may be a function of the well-controlled tasks and the quantitative nature of the judgment. Although we trained the judges not to simply look at the number correct within and across tasks but to consider the pattern of errors in arriving at their judgments, this may have been unavoidable.

One final note: Classical reliability estimates were high (and possibly were overestimated as indicated by the Rasch reliability estimates that account for extreme scores which inflate classical estimates). Given the nature of performance on these tasks, in which the correctness on an item is highly correlated with correctness on the task, this spuriously high level of reliability argues even more for the use of generalizability theory to more accurately estimate appropriate levels of variance attributable to specific sources.

Note

Support for this investigation was funded in part through an Enhanced Assessment Instruments grant (CFDA 84.368A – Grant # S368A040006) from the U.S. Department of Education, Office of Elementary and Secondary Education, awarded to the West Virginia Department of Education. The opinions expressed in this article do not necessarily reflect the position or policy of the Department of Education, and no official endorsement should be inferred.

References

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.). *Technical issues in performance assessments*. Washington DC: National Center for Education Statistics.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339-353.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: wAn example using conversational language samples. *Journal of Early Intervention*, *28*, 139-153.

Gao, X. & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*, 191-203.

Gierl, M. J. (1998). Generalizability of written-response scores for the Alberta Education English diploma examination. *The Alberta Journal of Educational Research*, *44*, 94-97.

Hintze, J. M., & Pettite, H. A. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment*, *19*, 158-170.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalzability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*, 71-92.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *Spring*, 5-8, 12.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Ruiz-Primo, A. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, *30*, 41-53.

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