Abstract

Critics of standardized testing have recommended replacing standardized tests with more authentic assessment measures, such as classroom assignments, projects, or portfolios rated by a panel of raters using common rubrics. Little research has examined the consistency of scores across multiple authentic assignments or the implications of this source of error on the generalizability of assessment results. This study provides a framework for conceptualizing measurement error when using authentic assessments and investigates the extent to which student writing performance may generalize across multiple tasks. Results from a generalizability study found that 77% of error variance may be attributable to differences within people across multiple writing assignments. Decision studies indicated that substantive improvements in reliability may be gained by increasing the number of assignments, as opposed to increasing the number of raters. Judgments about relative student performance may require closer scrutiny of task characteristics as a source of measurement error.

AUTHORS

John D. Hathcoat, Ph.D.
Oklahoma State University

Jeremy D. Penn, Ph.D.
Oklahoma State University

CORRESPONDENCE

Email
john.hathcoat@okstate.edu

Generalizability of Student Writing across Multiple Tasks: A Challenge for Authentic Assessment

For decades standardized testing in postsecondary education was limited to admissions testing. However, the influential report from the Secretary of Education's Commission on the Future of Higher Education recommended that all postsecondary education institutions should "measure student learning using quality-assessment data" with nationally standardized measures like the Collegiate Learning Assessment and institutions should make the results from those standardized tests "available to students and reported in the aggregate publicly" (United States Department of Education, 2006, p. 24). Critics have questioned the usefulness of standardized tests for both institutional accountability and institutional improvement. Common concerns with standardized tests include an overemphasis upon narrowly focused skills/abilities and content, the mismatch between the standardized tests and students' experiences at an institution, as well as students' motivation to complete such tests (Banta, 2006). Instead of standardized tests, researchers have suggested using what is called "authentic assessment," which includes approaches like assessment of e-portfolios, or assessment of writing and critical thinking (usually embedded in a course) using a common rubric (Banta, Griffin, Flateby, & Kahn, 2009).

Authentic assessment procedures may more directly reflect student experiences than standardized tests, though it remains unclear the extent to which it is appropriate to use authentic assessments in place of the many uses of standardized test scores. For instance, one desired use of standardized test scores is to compare students' performance across different schools (Benjamin, 2012). Standardized tests are standardized to control for specific sources of potential error – namely, differences in the characteristics of

If the assumption that student performance is consistent across multiple tasks is not tenable, and if authentic assessment is ever going to replace the numerous roles of standardized tests, then strategies must be developed to address task consistency.

tasks included within a test and the consistency of scores across alternative test forms. This does not negate many concerns with standardized tests, given that a standardized test score may reflect a single measure of a student's attribute, performance, or ability that fails to generalize to other settings. But authentic assessments, by their very nature, do not readily lend themselves to the same level of control across multiple sources of error. Just as score inconsistency across multiple items and/or alternate test forms restricts inferences from standardized tests, inferences derived from authentic assessments may be affected by multiple sources of error. Put differently, score consistency (i.e., reliability) is a necessary but insufficient condition to justify *any* use of scores deriving from an assessment regardless of whether it is standardized or authentic.

Examining the role of distinct sources of measurement error, along with interactions across these sources, remains of paramount importance in assessment practices. Such concerns, however, may lead to specific challenges for authentic assessment. It was therefore the purpose of the present study to address two concerns that that are potentially disconcerting for authentic assessment practices. First, common sampling strategies implicitly assume that some sources of measurement error are irrelevant. For example, many authentic assessment processes presume that *student performance is consistent across multiple tasks.* Users of assessment data may reasonably wonder if judgments about which students are doing best drastically changes across tasks or if measurement error is within acceptable limits. If the assumption that student performance is consistent across multiple tasks is not tenable, and if authentic assessment is ever going to replace the numerous roles of standardized tests, then strategies must be developed to address task consistency. Secondly, we investigate this assumption by applying generalizability theory to authentic assessment data (i.e., writing performance) collected at Oklahoma State University. Before proceeding to the study findings, a broad framework for vivisecting error variance through the lens of generalizability theory is provided.

Vivisecting Measurement Error across Multiple Tasks Using Generalizability Theory

This section provides an initial framework for conceptualizing the influence of assignment or task characteristics as a source of measurement error with respect to specific assessment goals and sampling strategies (Table 1). This framework is not meant to be inclusive, but is instead presented to illustrate a fundamental assumption with respect to sampling designs and measurement error: *If a single assessment or test is assumed to be representative of a student attribute, trait, or skill as a whole then evidence should be provided that such a use of that score is plausible.* This does not imply that assessment practitioners are explicitly aware of this principle when sampling specific assignments and/or tasks. In fact, we believe quite the contrary. In our own general education assessment practices at OSU, we have assumed that a single observation of a student's work is a reasonable estimate of performance when making comparisons at the institutional level. Although this is a low-stakes assessment for students, the kinds of inferences we hope to draw from this assessment process require that this assumption holds, and this may be specially concerning when employing specific sampling strategies (see Table 1). However, a failure to acknowledge or test this assumption does not render it unimportant. If this assumption is reasonable, judgments about student differences may be made irrespective of task characteristics. If this assumption is not tenable then judgments about students' performance may change if the researcher happened to sample a different task.

Evidence for person by task interaction effects may be particularly devastating given that this implies that judgments about which students are doing better depend upon the specific task that is assessed. With respect to writing assessment, this interaction would suggest that judgments about relative student differences drastically change across writing tasks or assignments, which may even occur within a single course. This particular source of measurement error can hinder assessment efforts targeting both within-group and between-group comparisons (Table 1). For example, two writing tasks, or assignments,

Table 1

Assessment Goals, Sampling Designs, and Measurement Error Related to Task Characteristics

Goals	Level of Generalization	Example Research Question	Sampling Design	Assignment as Source of Error
Between-Group / Cross-sectional	Institutional	On average are writing scores in 2012 higher than writing scores in 2011?	Random selection of student writing papers across level of interest.	Limited concerns; distribution of assignments should be checked across relevant comparisons.
	Program	Do students having experience 'X' tend to do better in writing than students who have not experienced 'X'?	Other than random	Potential bias due to distribution of assignments. Judgments about groups / individuals may change across assignments.
	Classroom	How well are students writing in this class?		
Within-Group	Institutional	On average, do freshman writing scores tend to improve by senior year?	Assignments have same prompt.	Limited concerns about assignment; check other issues stemming from design (e.g. practice effect)
	Program	How do writing scores change after participating in program 'X'?	Assignment has different prompts.	Potential bias resulting from a change in prompts.
	Classroom	Are writing scores improving across the semester?		Conduct G-study prior to large assessment, or design a study to control for prompt characteristics.

may be collected across the same students in order to assess changes in performance across time. Inferences about such changes are reasonable to the extent to which the two writing tasks are similar. A fundamental challenge, it would seem, is to provide evidence that tasks are sampled from the same theoretical domain, or the same universe of possible tasks. To once again place this argument within the context of writing assessment, claims about student writing performance must either be restricted to the specific task that is sampled, or evidence should be provided that performance generalizes across multiple tasks that are believed to be interchangeable.

If a single assessment or test is assumed to be representative of a student attribute, trait, or skill as a whole then evidence should be provided that such a use of that score is plausible.

Classical test theory (CTT), which is typically used to investigate score reliability via test-retest correlations, alternate forms, and/or internal consistency methods, is clearly limited for addressing these concerns. CTT, which assumes that an observed score may be decomposed into an expected true score and random error (Crocker & Algina, 1986), not only fails to consider multiple sources of error simultaneously but also fails to investigate interaction effects across sources of measurement error. Generalizability theory, or G-theory (Cronbach, Glesser, Nanda, & Rajaratman, 1972), has less restrictive assumptions than CTT and in many respects supplants this framework since it has been repeatedly demonstrated that investigations of reliability under CTT are special cases of G-theory designs (e.g., Brennan,

2011). Though both authentic assessment and G-theory have been utilized for some time now, for reasons that extend well beyond the scope of the present article, it appears that the utility of this approach for understanding sources of measurement error within the context of authentic assessment has yet to be fully realized. Others have addressed G-theory in detail (Brennan, 2001), and there are many good introductions to this topic (e.g., Shavelson & Webb, 1991). The following section will therefore close with a conceptual introduction to concepts employed within G-theory.

Conceptual Overview of Generalizability Theory

Claims about student writing performance must either be restricted to the specific task that is sampled, or evidence should be provided that performance generalizes across multiple tasks that are believed to be interchangeable.

G-theory utilizes analysis of variance techniques in order to further partition error into distinct sources of variation. These sources of variation are referred to as variance components, and estimating the relative magnitude of these components is of substantive interest in a *G-study*. A crucial task in designing a G-study is specifying the conditions of measurement, or *facets*, which presumably influence variation in observed scores. Facets may be either *crossed* or *nested*. A facet is considered crossed if every level of the first facet is observed at each level of the second facet (e.g., each student responds to every item), or alternatively a facet is considered nested within another if levels of one facet are observed at only one level of another facet (e.g., items may be nested within students if each student receives multiple items, but no student receives the same items). Facets may also be *random* or *fixed*. A facet is considered random if random sampling of each level has occurred or if the researcher is willing to treat observed levels as interchangeable (e.g., items may be replaceable with any other item of similar characteristics). A facet is considered fixed if the researcher has observed each level of facet or if the researcher does not wish to generalize beyond the observed levels of a facet.

Within a G-theory framework each observed level of a random facet may be viewed as a sample from a defined universe of acceptable observations. For example, within the context of writing assessment we are not necessarily interested in a student's performance on a specific assignment or writing task. *Instead, the specific task that is used may be viewed as one of many possible tasks that could have equally been utilized to assess writing performance*. In this case, we are interested in our ability to consistently estimate scores across tasks defining a universe of acceptable observations, irrespective of the specific writing task that was actually sampled in our assessment procedure. The generalizability coefficient ($E\rho^2$; Cronbach et al., 1972), which is the ratio of universe score variance to observed score variance (Webb, Shavelson, & Haertel, 2007), provides such an estimate by allowing us to examine the extent to which consistent estimates about relative student performance may be inferred across multiple tasks that are considered interchangeable. Generalizability coefficients range from 0 to 1.0, with acceptable coefficients ranging from .70 to .80 or higher (Brennan, 2001). Decision studies or D-studies may then be conducted to investigate how changes in specified facets may minimize error variance. We now summarize our own investigation of task variability as a source of measurement error within the context of general education assessment using G-theory.

Methods

Procedure

Each year Oklahoma State University (OSU) assesses the general education program and generates an annual report (http://tinyurl.com/osugened). This assessment effort typically involves sampling student papers (i.e., tasks) from courses across the campus. Each year tasks are sampled within the same semester, and faculty members act as paid raters who then score each paper in small independent groups of 2-3 members. Although the overall goal of the assessment process is to make general judgments about the extent to which students are achieving general education learning goals, as previously discussed, these judgments may still be affected by the task or assignment characteristics.

We began by examining the number of students for whom we had, by happenstance, scored more than one assignment or task in the entire set of data from 2001-2011. Of the scored areas, writing had been evaluated every year from 2001-2011, with the exception of 2007. In the 10 years in which writing was assessed there were a total of 1,831 scores, of which we identified 29 students who had more than one paper scored for writing. Of these, seven students had writing tasks sampled across different years of data collection. To avoid confounding results across years, these students were removed from subsequent analyses. The remaining 22 students were scored on two writing tasks sampled within the same semester, though each task was scored by an independent group of two faculty raters. This provided a total of 44 tasks, each of which was scored by two faculty raters, thus making 88 total observations. Since sample size may contribute to the stability of estimated variance components (Webb et al., 2007), the size of this design warrants some caution. However, the number of observations employed within this study is similar to many investigations utilizing G-theory.

Instead, the specific task that is used may be viewed as one of many possible tasks that could have equally been utilized to assess writing performance.

Instrumentation

Before faculty raters are assigned writing tasks to score they are first trained to use a rubric developed at OSU (see Appendix A). Scoring procedures have slightly varied throughout the years, though typically each faculty member rates tasks independently and then meets with their group in order to reach consensus with respect to each task's assigned score. Each task is scored on a 1-5 scale on content, organization, mechanics, and documentation so that higher scores reflect greater writing performance. In addition to dimensional scores faculty raters also provide an overall score reflective of the general writing performance exemplified by a student paper. The overall score provided by each faculty rater prior to consensus was utilized in the present analysis. Inter-rater reliability estimates tend to vary across groups of raters when approached under a CTT framework. A benefit of setting up a G-study is that distinct sources of error may be simultaneously examined in terms of their relative contribution to error. Reliability analyses are detailed in the results section.

Analytic Design

There were a total of 22 students who were sampled on two different writing tasks. Each task was scored by an independent group of two raters. Raters were therefore nested within tasks. However, given that each task was also different across persons tasks are considered to be nested within persons. Though there are statistical disadvantages to a fully nested design (i.e., confounded sources of error) this design resulted from restrictions deriving from decisions that were made about previous sampling strategies. Persons were treated as the object of measurement and both raters and tasks were conceptualized as a random sample from a potentially infinite number of observations. This entailed a fully nested, random effects design wherein the following variance components were estimated: persons (σ^2_P), tasks within persons $(\sigma^2_{T:P})$, and raters within tasks within persons $(\sigma^2_{R:T:P})$. The main effect of persons (σ^2_P) indicates the estimated variance component for between-person differences in average writing performance. Within the current study this variance component reflects the 'universe' from which we wish to make consistent inferences about student writing. The variance component for tasks within persons $(\sigma^2_{T:P})$ reflects mean differences in assignment scores for each person across the pairs of raters. Given that each task was assigned to a different group of raters this variance component cannot be disentangled from a person by task interaction. The variance component for raters nested within tasks nested within persons $(\sigma^2_{R:T:P})$ indicates differences in assigned scores within a single group of raters for a particular task. This source of variation is also confounded with unexamined sources of error.

Given that persons were the object of measurement we focused on the ability of scores to provide relative comparisons about inter-individual differences in writing

performance. In estimating the generalizability coefficient relative error is a function of both $(\sigma^2_{T:P})$ and $(\sigma^2_{R:T:P})$:

$$\sigma^2_{\delta} = \frac{\sigma^2_{T:P}}{n_T} + \frac{\sigma^2_{R:T:P}}{n_T n_R} \quad (1)$$

These values suggest that raters within each task tended to display little disagreement about the overall writing performance indicated by a particular student paper…A failure to find such differences however, indicates little about the consistency of rank ordering student writing ability.

From this equation it can be seen that both increases in $(\sigma^2_{T:P})$ and $(\sigma^2_{R:T:P})$ will inflate the amount of error in making normative judgments about student writing ability. Variation in average ratings assigned to tasks within a person and variation of raters within each task contribute to an inability to make consistent judgments about relative student writing performance. Increases in the number of observed tasks (n_T) and the number of assigned raters within each task (n_R) will decrease relative error given that the estimated variance components remain constant. Estimates of relative error are utilized in order to calculate the generalizability coefficient:

$$Gen\_Coefficient = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_{\delta}} \quad (2)$$

From equation 2 it can be seen that the generalizability coefficient is expressed as a ratio of total between person variation (i.e., universe score variance) to estimated observed score variation. Increases in the magnitude of relative (σ^2_{δ}) error will reduce the generalizability coefficient whereas increases in universe score variance (σ^2_P) will tend to increase the generalizability coefficient. As previously indicated this coefficient may be interpreted as the extent to which one may make consistent normative inferences about student writing performance across all possible raters and tasks.

Results

Descriptive statistics were first examined on the 22 students (i.e., 44 writing tasks) who had each task assessed by an independent group of two raters (see Table 2). It is of particular interest to note that the variation of assigned scores within raters for each task was relatively low. Within-task rater variance ranged from 0.00 to 1.00 with an average variance across each task of 0.13. These values suggest that raters within each task tended to display little disagreement about the overall writing performance indicated by a particular student paper. With such data many researchers may choose to utilize inferential statistics in order to investigate either mean difference across each writing task or the linear relationship between assigned scores across each writing task. For this analysis the mean rating provided by both judges for a single task was the outcome variable. A dependent sample t-test indicated no statistically significant differences across mean ratings assigned across writing tasks. A failure to find such differences however, indicates little about the consistency of rank ordering student writing ability. The observed correlation across each writing task was .178 (p = .429), which implies that the pattern of student writing scores across each task was highly inconsistent. Estimated variance components from the G-study were examined in order to investigate these inconsistencies using EduG 6.0 (Cardinet, Johnson, & Pini, 2010).

Results from this analysis are presented in Table 3. The object of measurement, between-person differences in student writing, consists of approximately 12% of the total variation. Though not large, this represents the signal that the assessment procedure is attempting to detect. Rater variation within each task that is also nested within each person consists of approximately 23% of the total error variation. Though the magnitude of this variation is substantive it is of particular interest that 77% of the error variance derives from differences within a single person across each task. As previously indicated, the design of this study confounds a task effect with a person by task interaction. The vast majority of error variance may be attributed to either a task effect or the possibility that the rank ordering of individuals in writing changes across each sample of tasks. The estimated generalizability coefficient was .28 (*SEM* = .55), which is far below acceptable standards. If we assume that error is normally distributed we may utilize the standard

Table 2

Task Variation and Mean across Raters

Person	Task	Task Mean	Task Variation
1	1	3.00	0.00
1	2	4.00	0.00
2	1	3.00	0.00
2	2	2.50	0.00
3	1	3.50	0.25
3	2	3.00	0.25
4	1	4.00	0.00
4	2	3.00	0.00
5	1	4.00	0.00
5	2	3.50	0.00
6	1	3.00	0.25
6	2	2.00	0.00
7	1	4.50	0.00
7	2	1.50	0.25
8	1	3.00	0.25
8	2	3.00	0.00
9	1	3.00	0.00
9	2	2.00	0.00
10	1	3.50	0.00
10	2	4.00	0.25
11	1	4.00	0.00
11	2	5.00	0.00
12	1	3.50	0.00
12	2	2.50	0.25
13	1	3.00	0.25
13	2	4.00	0.00
14	1	3.00	0.00
14	2	4.50	0.00
15	1	3.50	0.25
15	2	1.50	0.25
16	1	3.50	0.25
16	2	3.50	0.25
17	1	2.00	0.25
17	2	2.00	0.00
18	1	2.00	0.00
18	2	3.00	0.00
19	1	2.00	0.00
19	2	2.00	0.00
20	1	3.50	0.00
20	2	2.50	0.25
21	1	4.00	0.25
21	2	3.00	1.00
22	1	2.00	0.00
22	2	2.50	1.00

Note : Two raters are nested within each task. Tasks are also nested within students.

Table 3

Variance Component Estimates for Raters Nested in Tasks Nested in Persons Design

Source of Variance	*SS*	*df*	*MS*	Variance Estimate	Percent of Error Variance
Person	35.09	21	1.67	0.12	N/A
Task within Person	1.63	1	1.63	0.47	77.4%
Raters within Tasks within Persons	0.59	2	0.29	0.27	22.6%

Note: *SS* = sum of squares; *df* = degrees of freedom; *MS* = mean square. N/A = not applicable. N=22 persons rated by two groups of independent raters on two different tasks.

Examination of these confidence intervals suggest that individuals receiving a mean score of one…are indistinguishable from students assigned a mean score of two…though they may be distinguished from individuals assigned a score of three…or higher.

error of measurement in order to construct confidence intervals around mean scores on the writing rubric. Examination of these confidence intervals suggest that individuals receiving a mean score of one (95% CI = -0.078 to 2.078) are indistinguishable from students assigned a mean score of two (95% CI = 0.922 – 3.078), which in turn are indistinguishable from those with a mean score of three (95% CI = 2.922 – 4.078). Individuals with a mean score

It appears that, in order to use authentic assessments to make direct comparisons of students' scores, understanding the impact of task characteristics may very well be the biggest challenge.

of four (95% CI = 3.922 to 5.078) are, for all practical purposes, indistinguishable from students receiving a mean score of five (95% CI = 4.922 – 6.078). Stated differently, current assessment practices may distinguish those with relatively low scores (i.e., mean score of 1 and 2) from those with relatively high scores across both assignments (mean score of 4 and 5). However, more subtle distinctions in student performance across these tasks may not be consistently inferred.

Several D studies were conducted in order to evaluate the expected impact of increasing both the number of sampled writing tasks and the number of raters assigned to score each task. As indicated by Figure 1, little increase in the generalizability coefficient is predicted by increasing the number of raters assigned to each task. While holding the number of tasks constant at five the predicted generalizability coefficients range from .51 to .53 across three to seven raters. However, greater gains in the generalizability coefficient may be made from increasing the number of tasks collected on each person. While holding the number of raters constant at three the predicted generalizability coefficient ranges from .51 to .75 when increasing the number of tasks from 5 to 15. This pattern substantiates inferences from the G-study that suggested an increase in the number of raters assigned to a particular task may be of limited utility given the relative magnitude of error associated with differences assigned to tasks within a person. Instead, greater precision in making judgments about relative student writing performance may derive from increasing the number of observed writing tasks. Unfortunately, the number of tasks needed to substantially improve these inferences may be unobtainable in most assessment contexts due to the cost of collecting and scoring a substantially larger number of assignments.

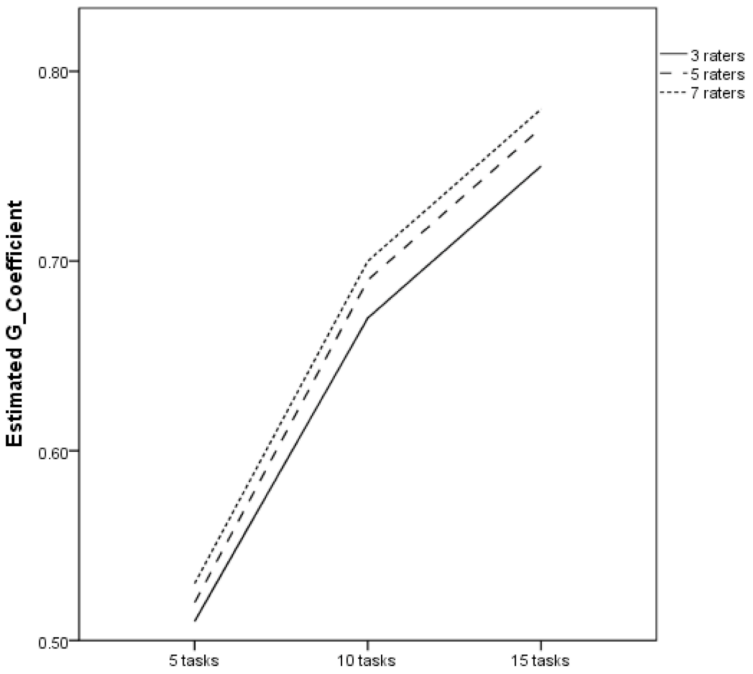

Figure 1. D studies for raters within each task by number of sampled tasks.

Discussion

Inter- and intra-rater reliability were once believed to be the biggest problems with authentic assessments. Instead, it appears that, in order to use authentic assessments to make direct comparisons of students' scores, understanding the impact of task characteristics may very well be the biggest challenge. Without an a priori equating of writing tasks, judgments derived from authentic assessment may largely depend upon the kinds of tasks students

are asked to perform. This is not to say that authentic assessment should be abandoned, nor does this evidence imply that standardized tests should replace authentic assessment. Instead, further investigation is needed to explicate the conditions under which generalized inferences are justified. The success of authentic assessment may therefore depend upon systematic efforts to articulate why judgments about relative student performance seems to change across separate tasks.

Within our sample, it was clear that the ordering of students by writing performance depended upon which task was selected. This study suggests that if researchers want to make comparisons about students' performance from authentic assessments between institutions or within an institution, they should greatly increase the number of tasks that are sampled for each student, establish statistical controls based on variables that are shown to impact students' performance (such as motivation), or take steps to standardize some task characteristics (which may not be palatable for users of authentic assessment). More than twenty years ago Elliot Eisner wrote, "Our nets define what we shall catch" (1992). Our study supports this statement by suggesting that what our students are able to show they can do is in part influenced by what we ask them to do.

If researchers want to make comparisons about students' performance from authentic assessments between institutions or within an institution, they should greatly increase the number of tasks that are sampled for each student, establish statistical controls based on variables that are shown to impact students' performance (such as motivation), or take steps to standardize some task characteristics.

While standardized tests may make a stronger case for controlling specific sources of measurement error, other aspects of standardized tests may not compare favorably with authentic assessments. First, if the content of a standardized test is selected at the national level and does not represent the goals, mission, and learning outcomes desired by an institution, it may be just as dubious to claim such a test is a reasonable comparison between institutions. Second, the extent to which scores on standardized tests extends to the kinds of tasks students perform throughout their education remains controversial. Just as our evidence implies that a single observation of writing performance fails to generalize to performance across other tasks, a similar issue may exist with standardized tests since these scores may also fail to generalize to scores observed on similar tasks outside a controlled testing environment. Third, there is some research to suggest that task characteristics are important to standardized tests as well. Russell and Plati's (2000) study illustrates this point. They found that student performance on a standardized test depended upon the mode of administration (computer or paper), and student performance was also a function of prior keyboarding skills. Even though standardized tests use a similar task across all examinees, the characteristics of the task still matter when making inferences using the scores from standardized tests.

Finally, regardless of whether authentic assessment or standardized tests are used to draw inferences, this study highlights the importance of explicitly addressing assumptions about the contribution of particular sources of measurement error. Specifically, when observing a single student assignment, or task, there are dangers in interpreting the scores as though they were independent of the task being sampled. Findings from the present study suggest interpretations that fail to account for task variation may be problematic since they presume that judgments about relative student differences are consistent across distinct tasks. Numerous authors have raised similar concerns (e.g., Kroll & Reid, 1994; Schoonen, 2005; Shavelson, Baxter, & Gao, 1993), and this study provides additional support that may serve to caution drawing unwarranted inferences from assessment results. Before proceeding to the implications of the present study, a central limitation will first be addressed.

Limitations

Numerous limitations exist with the present study; however, one limitation is particularly salient. Historical data were used in an effort to investigate the extent to which assumptions about the consistency of student performance across multiple tasks may be problematic. Methodological choices of previous assessment efforts restricted the analytic design employed within the current study. Within the current study raters were nested within writing tasks, which in turn were nested within persons. This design confounds important sources of error that may be important when deciding which strategies to adopt in subsequent assessment procedures. For example, this nested design makes it impossible

to disentangle a task effect and a person by task interaction. A fully crossed design would allow the separation of interaction effects between persons and tasks and persons and raters. Though the analytic design was not ideal, it provides tentative evidence in support of a growing concern about task variability as a source of measurement error within assessment practices.

Future Research

If a random number appeared with each observation of a pocket-watch, it would be challenging, but more importantly extremely misleading, to argue for the validity of a particular interpretation of these observed "times." We would not be able to use the pocket-watch to complete even a simple task accurately, such as putting students in order based on their time of arrival to class. No matter how carefully we analyze the scores from the random-number generating watch, they remain of little value. Without score consistency (e.g., we observe a similar time when each observation is conducted with the sun being at a particular point in the sky) nothing is being measured (Thompson, 2003). Reliability is thus a prerequisite, and principle justification, for the assignment of meaning to a set of scores. It is the role of validation to investigate evidential support for proposed interpretations given to a set of scores. Stated differently, reliability justifies the assignment of meaning to a set of scores, and validation is a constructive act whereby evidence is accumulated to articulate the limits, boundaries, and extension of a particular interpretation. Both reliability and validity are hence central considerations that inform decisions derived from educational assessment procedures.

When observing a single student assignment, or task, there are dangers in interpreting the scores as though they were independent of the task being sampled.

Fluctuations in student performance across multiple tasks, particularly if performance is a function of task characteristics, restrict the kinds of inferences that are justified when interpreting assessment results. Unfortunately, a simple panacea does not, at least as far as current research suggests, exist. As a first step, it is necessary to replicate the present findings in a study explicitly designed to control for confounded sources of error. Instead of using a nested design, a fully crossed design wherein every rater scored the same students on the same multiple tasks would be ideal from a G-theory perspective. Despite such constraints, the present results have led to concerns with our own assessment procedures, and additional data is currently being collected in an effort to further investigate the role of task characteristics as a source of measurement error when attempting to assess students' writing performance. Note that the present study is delimited to student writing performance, though we suspect that similar issues may arise when investigating other valued learning outcomes (e.g., critical thinking, intercultural competence, etc.). Examining this source of error in other authentic assessment processes (i.e., portfolios, critical thinking rubrics, etc.) is warranted. Though simple heuristic devices fail to account for the complexity of assessment practices, three general considerations are addressed that may be used to inform subsequent assessment efforts.

First, authentic assessment and G-theory have been discussed in the literature for some time now. Reliability estimation under classical test theory cannot address the complexity of task characteristics as a source of measurement error within authentic assessment practices. Consequently, we propose a "marriage" between G-theory and many assessment practices. G-theory provides greater flexibility to assessment practitioners, has less restrictive assumptions than classical test theory, and may be utilized to check data quality prior to implementing large scale investigations. Additionally, once specified, decision studies may be utilized to investigate ideal assessment procedures. The flexibility provided by G-theory does come at a cost. G-theory can be computationally complex and implementing this approach not only requires foresight into methodological design, but also careful consideration of how facets of measurement are specified as sources of measurement error. G-theory may not be appropriate for all assessment purposes, but continual advancement of this field appears to require practitioners to confront the challenges introduced by distinct sources of measurement error.

Second, person by task interaction effects may demand increased precision in how the universe of generalization is conceptualized. Stated differently, the presence of person by task interaction effects is suggestive of at least two possibilities that are in

need of subsequent investigation. First, it is conceivable that sampled writing tasks are interchangeable in that they are derived from the same theoretical domain. Under this view, one task should be equivalent to others in that the specific tasks that are sampled are indifferent with respect to judgments about relative student performance. The present analysis, which sampled two tasks, suggested that at least ten tasks may be necessary to derive reasonably consistent estimates about relative student differences. This could imply that the two sampled tasks, just by happenstance, failed to be representative of the universe of all possible tasks. Increasing the number of observations should therefore provide a better representation of the theoretical universe of all possible tasks.

An alternative interpretation is also possible. It is conceivable that the sampled tasks are not interchangeable, suggesting that it is mistaken to treat these tasks as a reflection of the same theoretical domain, or universe of generalization. In this case, either inferences about student writing must be restricted to the specific tasks that are sampled or greater care should be taken when conceptualizing the kinds of tasks that are judged to reflect the same theoretical domain. It is possible that writing tasks with characteristic "X" compose a separate universe of generalization than writing tasks with characteristic "Y." If so, then tasks may be sampled while controlling for characteristic "X," and consequently generalized inferences about relative student performance would be restricted to tasks denoted by such a characteristic. At this juncture there are many more questions than answers, and clearly more work is needed to investigate which of these alternatives may be more viable.

Reliability justifies the assignment of meaning to a set of scores, and validation is a constructive act whereby evidence is accumulated to articulate the limits, boundaries, and extension of a particular interpretation.

Finally, we wish to draw this discussion back to the controversies surrounding the issue of standardized tests and authentic assessment practices. As previously indicated, reliability estimates within authentic assessment practices, particularly with the use of rubrics, have generally focused on score consistency across or within raters (Finley, 2011/2012). Though controlling for this source of error remains important, this is only part of the story. Consistency across tasks is also an important source of error that stands in need of clarification. Elucidation of this source of measurement error, we contend, is intricately connected to criticisms of standardized tests, specifically criticisms residing in the question of whether *general* skills can be assessed (Banta & Pike, 2012; Benjamin, 2012). Person by task interaction effects, at least in principle, may be utilized as evidence to address such debates. For example, students may be given writing tasks across two disciplines that are then scored by trained raters using a common rubric. A person by task interaction would indicate that judgments about relative student differences changes across discipline, or in other words this evidence may suggest that performance is domain specific, which could then be used to argue for further refinement of the universe of generalization from which writing tasks are sampled. Alternatively, we could sample writing tasks within a single discipline utilizing the same procedures. A failure to find a person by task interaction may then imply that generalized inferences within a specific discipline are justified.

In conclusion, the current study suggests that caution is warranted when interpreting many assessment results. This caution stems from a generally unrecognized source of measurement error, namely the introduction of task variability. An accumulating body of evidence suggests that students' performance may be highly varied across tasks, and that judgments about which students are doing better may change across seemingly similar tasks. These problems can restrict warranted inferences from assessment results by limiting desired comparisons both within and between institutions. However, we do not universally reject authentic assessment as an important component of educational practice. To the contrary, we believe authentic assessment plays a critical role in evaluating educational programs and for making decisions about program improvement so long as such inferences carefully address distinct sources of measurement error investigated within this and other studies. This study underscores our concern with task variability as a source of measurement error, while acting as an invitation to other users of authentic and standardized assessment to join us in this investigation.

References

Banta, T. W. (2006). *A warning on measuring learning outcomes.* InsideHigherEd.com. Retrieved July 9, 2012 from http://www.insidehighered.com/views/2007/01/26/banta

Banta, T. W., Griffin, M., Flateby, T. L., & Kahn, S. (2009). *Three promising alternatives for assessing college students' knowledge and skills*. Occasional paper #2. National Institute for Learning Outcomes Assessment. Retrieved July 9, 2012 from http://learningoutcomesassessment.org/documents/AlternativesforAssessment.pdf

Banta, T.W., & Pike, G.R. (2012). *Making the case against—One more time. Occasional paper #15*. National Institute for Learning Outcomes Assessment. Retrieved July 9, 2012 from http://learningoutcomesassessment.org/documents/AlternativesforAssessment.pdf

Benjamin, R. (2012). The seven red herrings about standardized assessments in higher education. *National Institute for Learning Outcomes Assessment. Occasional paper #15*. National Institute for Learning Outcomes Assessment. Retrieved September 18, 2012 from http://learningoutcomesassessment.org/documents/HerringPaperFINAL.pdf

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

Brennan, R.L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1-21.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Taylor and Francis Group.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L.J., Glesser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: John Wiley.

Eisner, E. (1992). In Blekin, G.M., & Kelly, A.V. *Assessment in early childhood education*. United Kingdom: Sage.

Finley, A. P. (2011/2012). How reliable are the VALUE rubrics? *Peer Review*, *13*(4), 31-33.

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3*, 231-255.

Russell, M., & Plati, T. (2000). *Mode of administration effects on MCAS composition performance for grades four, eight, and ten*. Malden, MA: Massachusetts Department of Education.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1-30. doi: 0.1191/0265532205lt295oa

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

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Thompson, B.T. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-30). Thousand Oaks, CA: Sage.

United States Department of Education (2006). *A test of leadership: Charting the future of U.S. higher education* [Secretary of Education's Commission on the Future of Higher Education]. Washington, D.C.

Webb, N.M., Shavelson, R.J., & Haertel, E.H. (2007). Reliability coefficients and generalizability theory. In C.R. Rao and S. Sinharay (Eds.), *Handbook of Statistics, Vol. 26* (pp. 81-121). Oxford, UK: Elsevier B.V.

Appendix A

Writing Rubric Developed at Oklahoma State University

	Skill	Level of Achievement 1	2*	3	4**	5
A	Content	Topic is poorly developed; support is only vague or general; ideas are trite; wording is unclear, simplistic; reflects lack of understanding of topic and audience; minimally accomplishes goals of the assignment.		Topic is evident; some supporting detail; wording is generally clear; reflects understanding of topic and audience; generally accomplishes goals of the assignment.		Topic/thesis is clearly stated and well developed; details/wording is accurate, specific, appropriate for the topic & audience, with no digressions; evidence of effective, clear thinking; completely accomplishes the goals of the assignment.
B	Organization	Most paragraphs are rambling and unfocused; no clear beginning or ending paragraphs; inappropriate or missing sequence markers. No clear over-all organization		Most paragraphs are focused; discernible beginning and ending paragraphs; some appropriate sequence markers. Overall organization can be inferred and is appropriate for the assignment		Paragraphs are clearly focused and organized around a central theme; clear beginnings and ending paragraphs; appropriate, coherent sequences and sequence markers. Overall organization is clearly marked and is appropriate for the assignment
C	Style and Mechanics	Inappropriate or inaccurate word choice; repetitive words and sentence types; inappropriate or inconsistent point of view and tone. Frequent non-standard grammar, spelling, punctuation interferes with comprehension and writer's credibility.		Generally appropriate word choice; variety in vocabulary and sentence types; appropriate point of view and tone. Some non-standard grammar, spelling, and punctuation; errors do not generally interfere with comprehension or writer's credibility.		Word choice appropriate for the task; precise, vivid vocabulary; variety of sentence types; consistent and appropriate point of view and tone. Standard grammar, spelling, punctuation; no interference with comprehension or writer's credibility.
D	Documentation	In text and ending documentation are generally inconsistent and incomplete; cited information is not incorporated into the document.		In text and ending documentation are generally clear, consistent, and complete; cited information is somewhat incorporated into the document.		In text and ending documentation are clear, consistent, and complete; cited information is incorporated effectively into the document.