Studies in Applied Linguistics & TESOL at Teachers College, Columbia University, Vol. 21, No. 2, pp. 21–42
The Forum

Generalizability of Writing Scores and Language Program Placement Decisions: Score Dependability, Task Variability, and Score Profiles on an ESL Placement Test

Daniel Eskin
Teachers College, Columbia University

INTRODUCTION

Second Language (L2) testing has increasingly relied on performance assessment (e.g., a written essay, a spoken monologue) to evaluate "practical command of language acquired" (McNamara, 1996 as cited in Grabowski & Lin, 2019, p. 54). However, such forms of assessment entail more complex task design and subjective human scoring judgment (Bachman, 2004), raising challenges for score dependability and score use due to variability associated with task design (Deville & Chalhoub-Deville, 2006; Gebril, 2009; In'nami, & Koizumi, 2016; Schoonen, 2012), differences in rater behavior (Bachman, Lynch, & Mason, 1995; Lynch & McNamara, 1998), and rating rubric functionality, especially when consisting of multiple subscales (Grabowski, 2009; Grabowski & Lin, 2019; Sawaki, 2007, Xi, 2007).

For agencies that deliver high-stakes L2 proficiency exams (e.g., *Educational Testing Service,* ETS) a research agenda has been undertaken for years to examine the role of rater, task, and rubric as sources of variability into their performance assessments (Lee, 2006; Sawaki & Sinharay, 2013; Xi, 2007; Xi & Mollaun, 2006). However, these challenges are more acute for less-resourced institutions using L2 performance assessments for making score-based interpretations about test-takers, such as a language program making decisions based on placement test scores (Bachman, et. al, 1996; Sawaki & Xi, 2019; Vafaee & Yaghmaeyan, 2020). In such cases, the question becomes a veritable cost-benefit analysis, weighing the benefits of improving score dependability and practical constraints associated with time, cost, and resources (Bachman & Palmer, 1996).

Considering this challenge, the current study investigates test score variance and dependability on two performance assessment tasks from a writing section on an English language program placement exam. The organization that administers the exam, the *Community Language Program* (CLP), operated within Teacher College, Columbia University's in conjunction with its in Applied Linguistic and Teaching English to Speakers of Other Languages (TESOL), uses test scores from this exam to assign the test takers to a particular English as a Second Language (ESL) class at an appropriate proficiency level (i.e., beginner, intermediate, advanced).

The CLP caters to a diverse student population, consisting of adult learners of English of various cultural, professional, and educational backgrounds, and, moreover, a range L2 English proficiency levels, posing challenges for placing these learners at the appropriate level (Vafaee & Yaghmaeyan, 2020). The exam, comprised of six sections (i.e., meaning, grammar, listening, reading, speaking, and writing), yields composite scores for each subsection and for the entire

exam, the latter of which is used to place test takers in a given class level on the basis of pre-determined criteria (i.e., 'cut scores').

In this sense, the consistency in which the CLP placement test assigns a test taker to the appropriate level is of central concern since misclassification can have consequences for all stakeholders (i.e., CLP administrators, students, teachers) (Bachman, 2004; Bachman & Palmer, 1996; Vafaee & Yaghmaeyan, 2020). More specifically, the writing section and the potential sources of variability in the test scores produced from this section (e.g., task types, rater behavior, analytic rubric subscales) can bring to bear issues for making valid score-based inferences in terms of making placement decisions based on task-level, section-level, and test-level composite scores.

Given this importance, the current study uses the *Multivariate Generalizability Theory* (MG-theory) (Brennan, 2001; Webb, Shavelson, & Maddahian, 1983) to analyze test data from a single administration of the CLP placement test writing section for sources of score variability (e.g., task difficulty, rating severity, rubric functionality) and overall dependability. MG-theory is uniquely suited for modeling the relative effects of multiple sources of variance in written performance assessment and for providing evidence for the overall dependability of test scores for making classification decisions, such as placement at a given proficiency level (Sawaki, 2007).

In addition, the current study interprets the findings from MG-theory analysis to evaluate the claim that using test scores from the CLP placement test writing section for making placement decisions is appropriate based on an *Argument-based approach to Validation* (See Kane's, 2006 framework). Kane's (2006) framework, consisting of six inferences related to different stages of test design, development, and analysis, is meant to "provide a coherent means by which to assess the plausibility of the theoretical and empirical evidence used to support or refute the inferences and assumptions with a proposed test interpretation or use" (Purpura, 2011, p. 739). Specifically, this study attempts to provide empirical backing for supporting two particular inferences using MG-theory analysis: (1) the *Generalization* Inference, relating observed test score to an 'expected score' across different test conditions, and (2) the *Utilization* Inference, considering whether test score use is fair, meaningful, and appropriate for stakeholders in the testing process (e.g., test-taker, test-designer, test score user) (See overviews in Chapelle, 2011; Chapelle, 2008; Purpura, 2011; Xi, 2008).

As a backdrop to the analysis used in this study, a preliminary overview is provided on *Generalizability Theory* (G-theory), specifically, its unique suitability for modeling sources of variability (e.g., task variability) in performance assessment, and how it provides evidence for supporting absolute, classification-based placement decisions.

MG-theory and Performance Assessment

G-theory is a statistical model which can determine the extent that variance in observed test scores can be attributed to construct-relevant factors (i.e., test-taker ability), construct-irrelevant factors (e.g., rater severity), or factors moderating performance (e.g., task difficulty), an innovation over traditional measurements of *'test reliability'* in Classical Testing Theory (CTT) (Bachman, 2004; Shavelson & Webb, 1991; Crick & Brennan, 1983 as cited in Brennan, 2001). MG-theory is an extension on this model that can account for such factors when multiple measures are part of the testing context (Grabowski & Lin, 2019; Webb, Shavelson, & Maddahian, 1983).

One such application is the use of MG-theory to analyze analytic rubrics with multiple subscales (e.g., content, language, organization) in L2 performance assessment (Grabowski, 2009; Lee, 2006; Sato, 2011; Sawaki, 2007; Xi, 2007). Moreover, the statistical software used for MG-Theory, mGENOVA (v. 2.1, Brennan, 2001) provides information for evaluating the score variance and dependability within subscales of an analytic rubric, the interrelationships between subscales (i.e., 'Universe score correlations') and the relative contribution of each subscale to error variance and score dependability in a composite score (i.e., 'Effective weights') (Grabowski & Lin, 2019). To that end, the current study uses information from mGENOVA output for evaluating sources of desirable and undesirable score variance and dependability of the CLP placement test writing section.

MG-Theory and Modeling Variability in Performance Assessment

In terms of modeling variability associated with written and spoken performance, logically, the role the rater and rubric can play a role in such variability (See In'nami & Koizumi, 2015). However, the expectation generally is that the scoring procedures, and those responsible for executing those procedures, should remain consistent across test-taker, and that noticeable variability in rater behavior or rubric subscale functionality are sources of unexpected and, potentially, undesirable variability (Xi, 2007).

By contrast, evaluating role of *task* variability in score variance is, comparatively, more complex, requiring test developers to consider how the task have been designed tap the target construct (e.g., writing ability) and how they differ from one another (e.g., a request letter, a compare/contrast essay) (Deville & Chalhoub-Deville, 2006; Gebril, 2009; In'nami, & Koizumi, 2016; Schoonen, 2012; Xi & Mollaun, 2006). Simply put, "tasks differ and language learners might respond differently to task features" (Schoonen, 2012, p. 375). To consider such differences as 'construct-irrelevant' would be to overlook differential abilities across context inherent to any language learner, and how they may perform more or less *proficiently* across Target Language Use (TLU) domains (e.g., transactional domains, academic domains) (Deville & Chalhoub-Deville, 2006).

In light of this, researchers using MG-theory to model score variance in performance assessment have frequently observed the presence of variability associated with differentially difficult tasks (See In'nami & Koizumi, 2015 for a meta-analysis factors task variability L2 performance assessment). One study noted that it is "not uncommon in performance assessment especially in instances in which the tasks meant to tap different aspects of the same underlying construct" (Grabowski & Lin, 2019, p. 70). Moreover, a litany of 'contextual' task features have been found to influence the presence of task variability (e.g., type – independent task, integrated tasks, context – general or academic language use). For the CLP placement test writing section in question for the current study, the role of *differentially* difficult tasks, and the extent to which this difference in task difficulty is a construct-relevant source of score variability for making placement decisions in this context will be considered.

MG-Theory and Placement Decisions

The placement of a student in an ESL level by classifying them according to their test score against a pre-determined criteria (or cut-score) is by nature, an *absolute decision*. It is

informed primarily by two data-points, the test-takers score, the cut-score(s) for determining the classification, rather than their relative-standing compared to other test takers (Bachman, 2004).

An advantage of G-theory analysis is its ability to provide information for making *absolute* 'Criterion-Referenced' (CR) decisions for test taker classification (e.g., placement of student into an ESL level course based on performance above or below a 'cut score'), and *relative* 'Norm-referenced' (NR) decisions for rank-ordering test-takers by performance (e.g., selecting the top 10% candidates based on score) (Bachman, 2004). Since 'placement', as an institutional decision in an English language program, is typically an absolute CR decision (Bachman, Lynch & Mason, 1995; Sawaki, 2007), this paper will only refer to CR-related information from MG-theory analysis, the Dependability coefficient, 'phi' (φ), and absolute error variance (σ^2_{abs}).

With these considerations in mind, this paper examines two writing tasks of the CLP placement test writing section using MG-theory. To that end, performance on these tasks, scored through a rating procedure using an analytic rubric with three subscales, will be analyzed using MG-theory with an eye for differences in score variance and dependability across subscale, the interrelationships between the subscales, and the relative contribution of each subscale to composite score variance and dependability. The findings from these analyses will then be employed to provide empirical backing for the *Generalization* and *Utilization Inferences* in a Validity argument (Kane, 2006).

Research questions

The following research questions are addressed in this study, providing empirical backing for the abovementioned inferences.

1. What is the relative contribution of multiple sources of variation (i.e., test takers, task, and ratings) to total score variability on the writing section of the CLP placement test for each of the three subscales of the analytic rubric?
   (Empirical Backing for the *Generalization Inference*)
2. How dependable are the scores at the individual subscale level (i.e., Content Control, Organizational Control, Language Control) and at the composite score level for making absolute criterion-referenced placement decisions in the CLP placement test writing section? (Empirical Backing for the *Generalization Inference*)
3. To what extent are the components of writing ability (i.e., Content Control, Organizational Control, Language Control) related in the CLP placement test writing section? (Empirical Backing for the *Generalization Inference*)
4. To what degree does each analytic subscale for rating the CLP placement writing section tasks contribute to composite score universe-score variance?
   (Empirical Backing for the *Generalization Inference*)
5. Is it justifiable to combine individual analytic scales into a single composite on the CLP placement test writing section? (Empirical Backing for the *Utilization Inference*)
6. How many tasks and ratings would be optimal for reliably measuring writing ability in the CLP placement test writing section, while remaining practical to administer?
   (Empirical Backing for the *Utilization Inference*)

METHOD

To address the research questions mentioned above, this study used MG-theory analyses as a means of investigating test score variance and dependability on a writing section of a language placement test using two tasks and an analytic rubric with three subscales. The subsections below provide a description of the participants in the study, a description of the instruments and procedures utilized for scoring test performance, and a description of the analytic and statistical procedures used for analyzing the test data.

Participants and Context

The test-takers

Two hundred and eight adult learners ($n_p = 208$) participated in an English language placement test at the start of the Spring 2020 semester administered by the CLP. Based on the anonymized dataset available from this test administration, the demographics characteristics of the sample are not known. However, the CLP typically attracts adult learners from a diverse range of nationalities, age-groups, and professional and educational backgrounds (Vafaee & Yaghmaeyan, 2020).

The raters

To score performance on the writing section, nine doctoral students were recruited as raters from the Applied Linguistics and TESOL graduate program at Teachers College Columbia University ($n_r = 9$). The raters varied in their degree of English as a Second Language (ESL) teaching and testing experience. Prior to rating performance samples, the 'norming' process involved two stages. First, they were introduced to the writing tasks and the analytic rating rubric using an online rater training system. Following this, rating behavior using the rubric to score task responses was calibrated (i.e., normed) by using benchmarked sample responses with score rationales.

It should be noted that, due to the rating design used by the CLP placement test writing section as discussed in the *'Scoring Procedure'* subsection of this study, I will refer to 'ratings' (*r'*) rather in the multivariate G-theory analysis, rather than 'raters.'

Instrument

The Test

The CLP placement test battery consists of six sections: grammar, meaning, listening, speaking, reading, and writing. The writing section is strictly timed (45 minutes) and is comprised of two tasks, each targeting a different writing genre. Task 1 prompts test-takers to write a customer review about a retail experience. Task 2 prompts test-takers to take a position on an argument and support this position with reasoning. The instructions suggest that they use 15 minutes of the allotted 45 minutes for the first task and 30 minutes for the second task.

The Rubric

The construct of writing ability was operationalized using a scoring rubric for both tasks consisting of three subscales: (1) Content control, (2) Organizational control, and (3) Language control. Each observed variable of the rubric was rated on a six-point scale (0-5). Interestingly, the band descriptors for the Content control subscales differed slightly at the band levels of 3, 4 and 5, with a performance described according to differing genres. However, otherwise, the band descriptors at the lower end of the Content control scale were the same. Similarly, the band descriptors for the Organizational control scales, primarily focused on textual coherence, organization, and cohesion, and the Language control scales, mainly pertaining to lexico-grammatical range, accuracy, and precision, were the same for both tasks.

Data Collection Procedures

Test Administration

The CLP placement test writing section was delivered as part of the regular administration at the beginning of the Spring 2020 semester to place students who had signed up for courses into a given ESL class at a particular language proficiency level (i.e., Beginner, Intermediate, Advanced). The writing section, administered in person on computer, involved the test takers responding to two tasks.

Scoring Procedures

Following the test administration, each task response (i.e., 416 responses) was 'double marked', assigned ratings for each subscale of the analytic rubric on a 6-point scale (i.e., Content Control, Organizational Control, and Language Control) by two raters. In cases in which discrepancies in rating a response with an individual subscale was 2 points or more, a third rater was assigned to adjudicate that response.

To calculate a composite score for reporting writing section performance to test takers, the individual subscale ratings by each rater were added up to a total out 15 points (i.e., Rating 1 total, Rating 2 total) then averaged out of 15 points for both task 1 and 2. The resulting composite score used for scoring reporting, in this sense, was 'compensatory' in that stronger performance on one task could 'compensate' for weaker performance on another (Bachman, 2004, see Chapter 10).

Lastly, as noted above, the same two raters did *not* score the same task responses. For this reason, MG-theory can only consider the effects of 'ratings' rather than raters when using this design that this study has selected, thereby rendering the analysis of inter-rater reliability, or individual rater behavior impossible. For this reason, we will refer to any effects associated with ratings (*r'*) rather than raters (*r*) in this study.

Analyses

In order to investigate the variance and dependability of test scores across multiple subscales of an analytic rating rubric, MG-theory analysis was conducted among 208 test-taker

samples for both writing tasks, each of which were assigned two ratings using the statistical software, mGENOVA (v. 2.1), (cited in Brennan, 2001). To statistically model the relative effects of true test taker ability, task difficulty, rating severity, and their interactions and error, on overall test score variance, a two-step process is required in any G-theory analysis, consisting first of with a *Generalizability Study* (G-study), then a *Decision Study* (D-study) (Bachman, 2004; Brennan, 2001; Sawaki & Xi, 2019; Shavelson & Webb, 1991). To that end, the design required for modeling test score variance and dependability has been explained below for the CLP placement writing section.

The first stage in designing a G-study involves defining the *'Universe of Admissible Observations'*, the types of measurements treated as 'random', or interchangeable with one another (Bachman, 2004). For this study, the facets considered random and, thus, interchangeable are tasks (*t*) ($n_t = 2$) and ratings, not raters (See 'Scoring Procedure') (*r'*) ($n_{r'} = 2$) in that one task could admissibly switched with another, the same would be true for a rating with another. In G-Theory, test-takers or 'persons' (p) ($n_p = 208$) would be referred to as the object of measurement rather than a facet. In this dataset, the random facets are 'fully-crossed' in that both tasks have been rated twice for all two hundred and eight test takers. As such, the object of measurement and the random facets in this two-facet, fully crossed design are outlined in Table 1.

TABLE 1
Two-Facet Fully Crossed Design - Random Facets, Object of Measurement

Object of Measurement	Random Facets
Persons (*p*) ($n_p = 208$) – Test-taker Ability	Tasks (*t*) ($n_t = 2$) – Task Difficulty Ratings (*r'*) ($n_{r'} = 2$) – Rating Severity

Regarding the three subscales that comprised the analytic rubric, Content Control (CON), Organizational Control (ORG), and Language Control (LAN), in MG-theory, these would be treated as three 'levels of a fixed facet' ($n_v = 3$), which are not interchangeable with one another (Grabowski & Lin, 2019; Lee, 2006; Sato, 2011; Sawaki, 2007). This would allow for individual univariate G-theory analyses within each-subscale, and a multivariate G-theory analysis in terms of the inter-relationships between the rubric domains, and their relative contributions to composite score variance. The following MG-theory notation denotes a two-facet fully-crossed design with fixed facets: ($p^\bullet$ x $t^\bullet$ x $r'^\bullet$) with the filled circle in super-script (i.e.,$^\bullet$) representing that it is fully-crossed with all fixed facets (Grabowski & Lin, 2019). Table 2 presents the data setup for this design, serving as the structure for the mGENOVA control cards used for this analysis.

TABLE 2
Multivariate Two-Facet Fully Crossed Design with Three Fixed Facets

(fixed, υ)	**CON**				**ORG**				**LAN**			
(random)	T1		T2		T1		T2		T1		T2	
(random)	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
P1												
P208												

Next, the identification of different sources of score variability are identified for a G-study for a single observation. For this, a researcher needs to identify: (1) the object of measurement effect (persons – *p*), (2) the main random facet effects (tasks – *t*, ratings – *r'*), and (3) the interaction effects between each (i.e., 'persons by task' – *p x t*, 'persons by ratings' – *p x*

r', 'tasks by ratings' – *t x r'*, and the highest-order interaction, 'persons by tasks by ratings' plus error – *p x t x r'*, e). Computationally, we can delineate (i.e., *decompose*) the variance components within this two-facet, fully-crossed design across three levels of a fixed facet according to sources of variation in the test score variance, denoted as σ^2. Table 3 outlines the variance components by each level of the fixed facet.

TABLE 3
Two-Facet, Fully Crossed Design with Fixed Facets (Variance Components)

Source of Variation (i.e., 'Effect')	**CON**	**ORG**	**LAN**
Persons *(p)*	σ_p^2	σ_p^2	σ_p^2
Tasks (*t*)	σ_t^2	σ_t^2	σ_t^2
Ratings (*r'*)	$\sigma_{r'}^2$	$\sigma_{r'}^2$	$\sigma_{r'}^2$
Persons by tasks (*pt*)	σ_{pt}^2	σ_{pt}^2	σ_{pt}^2
Persons by ratings (*pr'*)	$\sigma_{pr'}^2$	$\sigma_{pr'}^2$	$\sigma_{pr'}^2$
Tasks by ratings (*tr'*)	$\sigma_{tr'}^2$	$\sigma_{tr'}^2$	$\sigma_{tr'}^2$
Persons by tasks by ratings, and error (*ptr', e*)	$\sigma_{ptr',e}^2$	$\sigma_{ptr',e}^2$	$\sigma_{ptr',e}^2$
Total	$\boldsymbol{\sigma^2(X_{ptr'})}$	$\boldsymbol{\sigma^2(X_{ptr'})}$	$\boldsymbol{\sigma^2(X_{ptr'})}$

CON= Content Control ORG= Organizational Control LAN = Language Control

The second stage in G-theory analysis, a D-study, then uses these variance components to model score variance and score dependability for the current measurement (e.g., $n_t = 2$, $n_{r'} = 2$), or other alternative measurement designs (Bachman, 2004). In MG-theory, both variance components *within* each fixed facet and covariance components *between* each fixed facet according to each source of variation are calculated. Additionally, mGENOVA uses this information from a given D-study to yield information pertaining to the interrelationships between each sub-scale (i.e., 'Universe Score Correlations') and the relative contributions of each sub-scale to error variance and universe score variance in the composite score (i.e., 'Effective Weights').

Lastly, to reiterate, since test scores on the CLP Placement writing section are used for making *absolute* classification decisions in the language program, MG theory analysis reports information only pertaining to error variance and score dependability for absolute criterion-referenced decisions (i.e., the Dependability coefficient 'phi', φ, and absolute error variance, σ^2_{abs}).

Considering the statistical model described above, MG-theory using mGENOVA, and CTT using SPSS, yields output relevant for addressing the following questions is provided in Table 4 below.

TABLE 4
Research Questions and Statistical output to examine
(Adapted from Grabowski & Lin, 2019, p. 63)

1. What is the relative contribution of multiple sources of variation (i.e., test takers, tasks, ratings, persons-by-tasks interaction, etc.) to test score variability on the writing section of the CLP placement test for each of the three subscales of the analytic rubric?
Relevant Output: - SPSS: A comparison of mean differences (means, standard deviations) by task and subscale, and independent *t*-tests to evaluate significance of mean differences across task and subscale. - mGENOVA: Variance Components estimates for multiple sources of variation for each subscale for multivariate D-study for current measurement design
2. How dependable are the scores at the individual subscale level (i.e., Content Control, Organizational Control, Language Control) and at the composite score level for making absolute criterion-referenced placement decisions in the CLP placement test writing section?
Relevant Output: - SPSS: Internal Consistency Reliability Analysis using Cronbach's alpha ($\boldsymbol{\alpha}$) for subscales of analytic rubric - mGENOVA: Dependability coefficients (φ) (for absolute decisions) for each subscale and composite score
3. To what extent are the components of writing ability (i.e., Content Control, Organizational Control, Language Control) related in the CLP placement test writing section?
Relevant Output: - SPSS: Pearson product-moment correlations between scores on each subscale of the rubric - mGENOVA: Universe-score correlations between each subscale of the rubric
4. To what degree does each analytic subscale for rating the CLP placement writing section tasks contribute to composite score universe-score variance?
Relevant Output: - mGENOVA: Effective weights of the individual subscales of the analytic rating rubric for writing tasks to overall composite writing score on the CLP Placement test writing section
5. Is it justifiable to combine individual analytic scales into a single composite on the CLP placement test writing section?
Relevant Output: - mGENOVA: Effective Weights of individual subscales & Universe-score correlations
6. How many tasks and ratings would be optimal for reliably measuring writing ability in the CLP placement test writing section, while still remaining practical to administer?
Relevant Output: - mGENOVA: Dependability coefficients for each subscale and the composite for multivariate D-studies conducted for alternative measurement design conditions for tasks and ratings.

RESULTS

The following section reports the results from SPSS- and mGENOVA-derived output in order to answer the six research questions mentioned above in Table 4.

Research Question 1

To address the first research question regarding the relative contribution of multiple sources of variation (i.e., test-takers, tasks, ratings, persons-by-tasks interaction, etc.) to total score variability on the CLP placement writing section, two relevant data outputs were examined. The first, a comparison of mean differences by subscale and by task (Table 5), was computed in SPSS using composite averages of rubric ratings (out of 5 points) and composite writing averages of rubric ratings (out of 5 points). These mean differences were then submitted to independent *t*-tests using SPSS. Second, the variance components attributable to a particular source of variation (e.g., persons, tasks, ratings, persons-by-tasks interaction) by subscale was computed in mGENOVA, then the percentage of variance explaining total variance within that subscale was computed in Excel (Table 6). For reference, the dependability coefficient (φ) and absolute error variance (σ^2_{abs}) are included in Table 6, but not interpreted.

Considering the average ratings by subscale (i.e., Content, Organization, Language Control) and by task (i.e., Task 1 - customer review, Task 2 – argumentative essay), provided in Table 5, patterns in the mean differences between subscales and between tasks have been reported to describe performance overall, but also to identify potential sources of variation that could manifest in subsequent MG-theory analysis. To that end, each mean difference across tasks was submitted to independent *t*-tests below.

TABLE 5
A Comparison of Means by Writing Task and Subscale

	Task 1		**Task 2**		**Independent *t*-test**	
	M	SD	*M*	SD	*t*(414)	*p*
Subscale Average*	**3.14 out of 5**	**1.01**	**2.76 out of 5**	**1.04**	**3.780**	**.000**
Content Control	3.35 out of 5	1.11	2.85 out of 5	1.12	4.573	.000
Organizational Control	3.02 out of 5	1.06	2.64 out of 5	1.10	3.588	.000
Language Control	3.05 out of 5	0.98	2.78 out of 5	1.07	2.684	.008

**Averaged ratings across three subscales out of 5 points*

As can be seen from Table 5 above, a clear pattern emerges across task and subscale. In terms of the mean differences across task, Task 1 demonstrated higher mean ratings than Task 2 on averages of all the subscales (*M = 3.14, SD = 1.01* for Task 1 and *M = 2.76, SD = 1.04*). This mean difference across task was submitted to an independent *t*-test. The results indicate a statistically significant difference in means between the two tasks, *M = 0.38,* 95% CI [0.18, 0.57], $t(414) = 3.780$, $p < .001$. Based on these results, one can conclude that Task 2, the argumentative essay, was interpreted as a more difficult task for test takers than Task 1, the

customer review, or alternatively, that the raters interpreted the rating rubric more severely in the case of Task 2 than for Task 1.

A similar pattern emerged within each subscale when comparing the tasks. Content Control scale revealed the highest mean ratings regardless of task (M = 3.35, SD = *1.11* for Task 1 and M = 2.85, SD = *1.12* for Task 2), but when compared to one another using an independent t-test, revealed the most prominent and statistically significant differences, M = *0.50*, 95% CI [0.29, 0.71], $t(414) = 4.573$, $p < .001$. The Organizational Control scale appeared to be rated most severely regardless of task (M = 3.02, SD = *1.06,* for Task 1 and M = 2.64, SD = *1.10* for Task 2), but, likewise, revealed significant differences across task, M = *0.38*, 95% CI [0.17, 0.58], $t(414) = 3.588$, $p < .001$. Lastly, the Language Control scale showed the least prominent mean difference across task (M = 3.05, SD = *0.98* for Task 1 and M = 2.78, SD = 1.07 for Task 2), but when submitted to an independent t-test still revealed a statistically significant difference, albeit not quite as pronounced as with the other subscales, M = 0.27, 95% CI [0.07, 0.47], $t(414) = 2.684$, $p = .008$.

What these findings indicate is that the Content Control scale was interpreted more leniently on both tasks compared to the Organization Control scale and Language Control scale. However, almost systematically, within each subscale, performance on the customer review task (i.e., Task 1) was rated noticeably more leniently than the argumentative essay task (i.e., Task 2), or said another way, the latter more severely than the former, for content, organization, and language, indicating differences either in actual written performance, rater perceptions about written performance, or rater perceptions about how the rubric subscales *should* be applied to written performance.

So, what does this mean for subsequent MG-theory analysis? What is clear is that there are noticeable *mean differences* across subscale and across tasks. For the latter, such differences would typically manifest in the 'task' main (T) effect when analyzing variance components in a D-study. It also should be noted that the 'persons-by-tasks' interaction ($p \times T$) effect, while a measure of variance attributed to differences in how test takers are rank-ordered differently by tasks, rather than *mean* differences in the task difficulty (Grabowski, 2009; Grabowski & Lin, 2019; Sawaki & Xi, 2019; Vafaee & Yaghmaeyan, 2020), the presence of remarkably lower mean rating for one task over another may also potentially be influenced by *differentially difficult* tasks (Schoonen, 2012; Xi & Mollaun, 2006) in the event that test-takers who performed the best on the customer review did not perform the best on the argumentative essay, and similar for those who performed the worst on each task.

As an empirical question, the information yielded from MG-theory analysis will allow us to determine this the extent to which the customer review task and the argumentative essay task are *differentially difficult* by interpreting the persons-by-task interaction using MG-Theory analyses, but by comparison, using the information yielded from CTT analyses (e.g., descriptive statistics, independent t-tests), we can only speculate on whether these tasks are, in fact, differentially difficult (Brennan, 2000).

With these considerations in mind, Table 6 presents the variance components associated with each source of variation across the three levels of the fixed facet (i.e., CON, ORG, LAN), and the percentage of total variance within that subscale explained by a given 'effect' (i.e., source of variation). Particular attention has been paid to the following sources of variation: (1) the percentage of total variance within a subscale attributed to the 'persons'(p) effect, or *mean* differences in test taker ability, also known as the "Universe Score" and associated with construct-relevant variance in a given test (Brennan, 2001; Bachman, 2004; Shavelson & Webb,

1991), (2) the percentage of total variance within a subscale attributable to the 'task'(*T*) effect, or *mean* differences in task difficulty, assumed to be present across subscale given the results from Table 5, (3) the percentage of total variance within a subscale ascribed to the 'persons-by-tasks interaction' (*p* x *T*) effect, indicating differences in how the tasks *rank-order* test takers, an effect that has shown to be quite common in performance assessment (See In'nami & Koizumi, 2015 for a meta-analysis factors task variability L2 performance assessment), and (4) the relative absence of total variance explained by 'ratings' effect (*R'*) or any interaction involving ratings (*p x R'*, *T x R'*).

As noted previously, the dependability coefficient (φ) and absolute error variance (σ^2_{abs}) have also been reported in Table 6, but will be discussed in greater detail in reference to the second research question.

TABLE 6
D-study Variance Components, Error, and Phi by Scale (2 tasks, 2 ratings)

	CON		**ORG**		**LAN**	
Source of variation	σ^2	% var	σ^2	% var	σ^2	% var
Persons (*p*)	0.76259	**71**	0.71529	**73**	0.74070	**79**
Tasks (*T*)	0.06105	**6**	0.03264	**3**	0.01670	**2**
Ratings (*R'*)	0.00592	**1**	0.00000	**0**	0.00499	**1**
pT	0.16251	**15**	0.14043	**14**	0.08546	**9**
pR'	0.02413	**2**	0.02019	**2**	0.04069	**4**
TR'	0.00000	**0**	0.00104	**0**	0.00013	**0**
pTR',e	0.06484	**6**	0.07197	**7**	0.05035	**5**
Total	**1.08104**	**100**	**0.98156**	**100**	**0.93902**	**100**
Absolute error variance (σ^2_{abs})	0.31845		0.26629		0.19832	
Dependability coefficient Phi (φ)	0.71		0.73		0.79	

CON= Content Control ORG= Organizational Control LAN = Language Control

As described above, the total variance explained by 'persons' (*p*) in G-theory analysis is interpreted as true *mean* differences in test-taker ability, and can be considered a direct reflection of score dependability, referred to as the 'Universe score' for that given subscale. In terms of what levels of a 'persons' effect would be considered *dependable*, researchers in L2 assessment have come to generally agree that, "large variance component estimates (i.e., larger than 75%, say, for low stakes decisions, and larger than 90% for high-stakes decisions) for the persons effect are desirable because they indicate true differences in test taker ability" (Grabowski & Lin, 2019, p. 68).

With this in mind, and considering the CLP placement test, and the writing section tasks that comprise it, as a somewhat *lower* stakes context than high-stakes L2 proficiency exams (e.g., TOEFL iBT), we can consider the Language Control scale as performing most dependably, with 79% of variance within this subscale explained by true mean differences in terms of test taker

ability, followed by Organizational Control (73%), and Content Control (71%) performing least dependably, which is interesting, given the observably higher mean ratings yielded across task from this subscale (see Table 5).

Regarding the 'task' (*T*) main effect, which can be interpreted as variance explained by *mean* differences in task difficulty and was expected to explain some degree of total variance across subscale given the noticeably different mean ratings for task 1 and task 2, the presence of this effect ranged from 2% for the Language control scale, 3% for the Organizational control scale, and 6% for Content control scale.

These findings on the 'task' main effect are, in fact, supported by the task-level means for each scale (see Table 5), with Content Control (i.e., 3.35 - 2.85 = 0.50 out 5 point difference in task means) showing the widest gap between task means, followed by the Organizational Control scale (i.e., 3.02 - 2.64 = 0.38 out 5 point difference in task means), and the Language Control with the smallest gap between task means (i.e., 3.05 – 2.78 = 0.27 out of 5 point difference in task means). Moreover, when submitted to independent *t*-tests (see Table 5), all mean differences were found to be significant within subscale across task, but most prominently in Content Control, followed by Organizational Control, and least prominently in Language Control. What this suggests is that despite being the most leniently rated scale, raters interpreted the Content Control scale more severely for the argumentative essay (Task 2), and more leniently for the customer review (Task 1), revealing a considerable degree of score variability across task when rated for content.

In regards to the 'persons-by-task' interaction (*p* x *T*) effect, conveying differences in how test takers are rank-ordered by task in terms of ability, a similar pattern emerged as the 'task' (*T*) main effect. Once again, the Content Control scale revealed the greatest variance associated with this effect (15%), followed by the Organizational Control scale (14%), and the Language Control scale (9%) with the least variance explained by this interaction.

Taken together with the considerable *mean* differences observed between tasks across subscale, but especially for Content Control, we can assume that the two tasks were *differentially difficult* for test takers in all cases (Grabowski, 2009; Schoonen, 2012; Xi & Mollaun, 2006). Simply put, the rank-ordering of test takers from best to worst for the customer review was not the same the rank-ordering of test takers from best to worst for the argumentative essay for all three subscales. However, as evidenced by the *p x t* interaction being largest for Content Control, this lack of consistency in the two tasks rank ordering test takers was most present when the content of each task was rated.

Given that content is task-specific, and these differing tasks do still both tap the construct of language ability, albeit in different ways, I will discuss in the conclusion of this study the extent to which this large *p x t* interaction for content is expected, and in fact, construct-relevant given the differing writing abilities that the two tasks on the CLP placement test writing section tap (Deville & Chalhoub-Deville, 2006).

Lastly, the lack of total variance explained by the main 'ratings' effect (*R'*) across subscale (i.e., CON, 1%, ORG, 0%, LAN, 1%), the relatively small percentage of variance explained by the 'persons-by-ratings' interaction effect across subscale (i.e., CON, 2%, ORG, 2%, LAN, 4%) indicate, in the case of the former, little to no mean differences in rating severity, and relatively minor differences in test takers in terms of ability being rank-ordered differently by ratings in terms of severity.

Research Question 2

To address the second question, on the dependability of scores within each subscale and for the composite score for the CLP placement test writing section, two relevant output were examined: (1) from SPSS, Internal Consistency Reliability analysis using Cronbach's alpha (α), a measure of the homogeneity (i.e., consistency) in how the subscales of an analytic rubric perform (Bachman, 2004; Carr, 2010), and (2) from mGENOVA, the dependability coefficient, phi (\) by subscale and the composite, an analogue of CTT's reliability coefficient (α), but specifically for absolute decisions in G-theory (Bachman, 2004). It also should be noted that the 'phi' coefficient serves as an analogous calculation to the 'Universe score' by subscale (i.e., the percentage of variance explained by 'persons') (see Table 6).

Table 7 presents the internal consistency reliability between the three subscales using Cronbach's alpha (α) for task 1, task 2, and for the entire writing section.

TABLE 7
Internal Consistency Reliability for CLP Placement Test - Writing Section

	Writing Task 1	Writing Task 2	**Writing Section Average**
Cronbach's Alpha (α)	.96	.95	**.97**
N of subscales	3	3	**3**

The picture that emerges from these measures of internal consistency reliability is that, despite the sources of variance not explained by 'Persons' in MG-theory analysis, the three subscales, in fact, perform quite consistently. For even high-stakes L2 testing, a commonly referenced threshold for acceptable reliability is .90 (Carr, 2010), suggesting the reliability estimates for task (.95, .96) and the writing section (.97) are, beyond acceptable, and in fact, quite desirable.

In light of these remarkably high reliability estimates using CTT indices, the dependability coefficient, phi (φ), paint a more complex picture regarding the subscales, indicating a much lower reliability estimates, if we consider the 'phi' in G-theory as an analogue to the 'alpha' in CTT (Bachman, 2004). Table 8 presents the dependability coefficients by subscale and for the overall composite score.

TABLE 8
D-study Score Dependability by scale and composite (2 tasks, 2 ratings)

	Absolute error variance (σ^2_{abs})	Dependability coefficient Phi (φ)
Content control	0.31845	0.71
Organizational control	0.26629	0.73
Language control	0.19832	0.79
Composite	**0.20379**	**0.78**

As can be seen, when considered by subscale and composite, the phi estimates ranged from .71 to .79 by subscale, indicating the lowest dependability for the Content control scale (.71), followed by the Organization control scale (.73) and the highest dependability for

Language control scale (.79), yielding a dependability coefficient for the overall writing section composite score of .78.

When interpreted through the previously reviewed variance components by scale (see Table 6), it is fair to conclude the relative effects of mean differences in task difficulty (i.e., 'tasks', *t*), and rank-ordering differences of persons in terms of ability by tasks in terms of difficulty (i.e., 'persons-by-tasks' interaction, *p* x *t*), which were more prominent in the Content control scale and Organizational control scale, likely led to lower score dependability within these scales. These sources of variance pose an added layer of complexity for MG-Theory analyses that will be discussed in the following sections that aggregate the modeling of variance components across tasks (i.e., Universe-Score Correlations, Effective Weights). With that said, contemporary conceptualizations of Construct Validity (Deville & Chalhoub-Deville, 2006) and moreover, those using MG-Theory for making validity claims (e.g., Chapelle, 2008; Grabowski & Lin, 2019; In'nami & Koizumi, 2016) treat the presence of task effect and persons-by-task interaction effect as not only common in L2 written performance assessment but also relevant to the construct of writing ability.

Research Question 3

To investigate the extent that the components of writing ability on the CLP placement test writing section (i.e., Content control, Organizational control, and Language control) are interrelated, two different correlational estimates were computed: (1) Universe-score correlations, using mGENOVA and based on covariance estimates between the subscales for variance attributed to the 'Persons' effect (Webb, Shavelson, & Maddahian, 1983), and (2) Pearson product-moment correlations, using SPSS and based on covariance between average test taker scores within each scale. The two coefficients differ in that Universe-score correlations are disattenuated from measurement error (i.e., only construct-relevant variance considered, *'persons'*) while Pearson product-moments are not disattenuated from error (Grabowski & Lin, 2019; Xi, 2007), causing the Pearson correlational estimates to, potentially, be "artificially lower than (what) the true relationship embodies" (Grabowski, 2009, p. 160).

With this in mind, the correlation matrix in Table 9 presents the estimates of the interrelationships between the subscales of the analytic rubric using Universe-score correlations (in bold-type) and Pearson product-moment correlations (in parentheses).

TABLE 9
D-study Universe score correlations and Pearson product-moment correlations among the Three analytic rating scales (2 tasks, 2 ratings)**

Rating scale	CON	ORG	LAN
Content Control (CON)			
Organizational Control (ORG)	**1.01** (0.94*)		
Language Control (LAN)	**0.98** (0.88*)	**1.00** (0.92*)	

*Statistically sig. at the .01 level

**Universe score correlations are in bold; Pearson product-moment correlations are in parentheses

As seen above, the Universe-score correlations and Pearson product-moment correlations both revealed highly interrelated components of writing ability on the test, with the former ranging from .98 to 1.01, and the latter ranging from .88 to .94, all of which were statistically significant to a .01 level, meaning that there was a 99% probability that the Pearson product-moment correlations were not due to chance. Based on these estimates, the most inter-related scales were Content Control and Organizational Control (1.01, .94*) with slightly less, though still extremely high inter-relationships between Organizational Control and Language Control (1.00, .92*) and Content Control and Language Control (.98, .88*). Admittedly, Universe-Score Correlations over 1.00 may, potentially, suggest the presence of a 'hidden facet' not accounted for in the measurement (Brennan, 2001 as cited in Grabowski, 2009, p. 162). With that said, to find theoretical support for this observed overlap between the scales, one would simply need refer to the models of *Communicative Language Ability* (CLA) that have become well-accepted in the field of L2 assessment (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980) accounting for the role of propositional content, organization and rhetoric, and language (e.g. vocabulary and grammar) as constituent components in written and spoken language ability.

It should be noted that the estimates given very little evidence regarding the 'distinctness' of subscales, which could be considered important in terms of the proper functioning of the subscales (Sawaki, 2007). Nonetheless, on balance, these highly interrelated scales of the analytic rubric can be regarded as unsurprising and fairly well-supported by theoretical models delineating aspects of CLA. Such interrelatedness of scales can serve as the basis for moving to a holistic rating scale for Writing Ability. However, in doing so, the CLP would lose the opportunity to collect subscale-level data so useful for research purposes such as this very study.

Research Question 4

To examine the relative contributions of each subscale to composite score variance, the 'effective weights' by subscale to composite Universe-score variance and Absolute Error variance were computed in mGENOVA and are compared against the *a priori* nominal weight of 33.3% for each subscale in Table 10.

Table 10. **Relative contributions of sub-scale to composite universe score variance**

	CON	ORG	LAN
Nominal Weight	33.3%	33.3%	33.3%
Effective Weights			
Contributions to…			
Composite Universe Score Variance	34%	33%	33%
Composite Absolute Error Variance	36%	35%	28%

CON= Content Control ORG= Organizational Control LAN = Language Control

The results of this analysis can be considered largely unremarkable in the case of the relative contributions to composite Universe-score variance. Despite any differences in score dependability with each subscale (see Table 8), these distinctions at the subscale level did not manifest in any meaningful difference in the 'effective weights' of each scale for composite score variance, remaining almost identical to the nominal weights (i.e., ranging from 33% to 34%).

Interestingly, a more salient pattern did, however, emerge in the case of the relative contribution of subscale to composite Absolute error variance. Here, we can consider these effective weights as operating in conjunction with dependability by scale, with Language control, the most dependable scale (φ = .79) forming the smallest contribution to absolute error in the composite (28%), and the less dependable scales, Organizational control (φ = .73) and Content control (φ = .71) forming comparatively larger contributions to absolute error in the composite (35% and 36%, respectively). However, these results, while interesting, do not constitute a surprising finding that diverged considerably from the pre-established equal weighting scheme in the CLP placement test writing section.

Research Question 5

The findings from the previous two research questions suggest (1) the high interrelationships between the components of writing ability on the CLP placement test (e.g. Universe-score correlations, Pearson product-moment correlations), and (2) similar relative contributions of each subscale to composite score variance (e.g., 'Effective weights'). Taken together, one way to interpret these findings is that "the four scales are measuring the same underlying ability" (Vafaee & Yaghmaeyan, 2020, p. 93), that is, writing ability for this study. In terms of informing test score use, one can use the high related, yet equally contributing scales to the composite as a justification for the current CLP practice combining the subscales into a composite writing section score (i.e., out 15 points, see 'Scoring procedures' and Table 5 above), and reporting that score to test takers (Vafaee & Yaghmaeyan, 2020) since the findings suggest each component to be representative of test taker writing ability. Admittedly, this justification of combining subscale scores could be, potentially, undermined by the potentially meaningful *task-level* differences in scores, which are 'compensated' when combined in a composite (Bachman, 2004; Sawaki, 2007). For this reason, we will discuss the possibility of reporting task-level composite scores to test takers, in the next section, the 'Discussion'.

Research Question 6

To evaluate the impact that increasing or decreasing the number of tasks and ratings would have on score dependability within each subscale and for the overall composite, a series of D-studies measuring alternative measurement designs were conducted in mGENOVA for the following conditions: (1) varying the number of tasks but keeping ratings constant ($n_{r'}$=2), and (2) varying the number of ratings but keeping task constant (n_t=2). The dependability coefficients across subscale and composite for each condition, and the change in "phi" (φ) for the composite is provided in Table 11 and Table 12.

TABLE 11
Changes in dependability coefficient when varying the number of tasks

Number of tasks	CON	ORG	LAN	**Composite**	**Change in Phi (φ)**
1	0.56	0.62	0.68	**0.67**	**-0.11**
2*	0.71	0.73	0.79	**0.78**	**--**
3	0.77	0.80	0.83	**0.82**	**+0.04**

CON= Content control ORG= Organizational control LAN = Language control
**Current operational design (2 tasks, 2 ratings)*

TABLE 12
Changes in dependability coefficient when varying the number of ratings

Number of ratings	CON	ORG	LAN	**Composite**	**Change in Phi (φ)**
1	0.65	0.67	0.72	**0.71**	**-0.07**
2*	0.71	0.73	0.79	**0.78**	**--**
3	0.73	0.75	0.82	**0.81**	**+0.03**

CON= Content control ORG= Organizational control LAN = Language control
**Current operational design (2 tasks, 2 ratings)*

Overall, the results of the alternative measurement D-studies reflect a rather predictable outcome, especially when considered in tandem with the variance components within each subscale for the current measurement design (see Table 6). To illustrate, likely due to greater variability associated with 'tasks' (*T*) and the 'persons-by-tasks' interaction (*p x T)* effects in comparison with 'ratings' (*R'*) and the 'persons-by-ratings' intereaction (*p x R'*) effects across subscale, adjusting the conditions for 'tasks' had a larger influence on dependability ($n_t = 1$, Change in φ = -.11, $n_t = 3$, Change in φ = +.04) than adjusting the conditions for 'ratings' ($n_{r'} = 1$, Change in φ = -.07, $n_{r'} = 3$, Change in φ = +.03).

However, for language schools with relatively limited resources for administering and scoring their placement tests, such as the CLP, these considerations of increasing dependability must be balanced with practical constraints of time and human resources (e.g., the number of available raters recruited to score written performance tasks, the number of test administrations to administer the test) (Bachman & Palmer, 1996). For this reason, we will consider these findings as a justification for keeping the current measurement design ($n_t = 2$, $n_{r'} = 2$).

DISCUSSION

The purpose of this study was to examine test score variability and dependability on the writing section one language program's ESL placement. Using the Kane's (2006) Argument-based approach to Validation (Chapelle, 2008; Chapelle, 2011; Purpura, 2011; Xi, 2008), the findings from this study have been considered in terms of the *'Generalization Inference'* linking observed test scores to expected scores across test context when accounting for the dependability of observed test scores, and the *'Utilization Inference'*, linking test score interpretations to test score use for stakeholders (e.g., the school, the students, the teachers).

As such, the findings from the first four research question provide empirical evidence that can be considered for backing claims meant to support the *'Generalization Inference'* linking test

scores from this CLP placement test administration to expected scores when considering their dependability across test administration (Purpura, 2011). In terms of dependability, the findings using the traditional CTT approach to reliability (i.e., Cronbach's alpha - α) displayed remarkable levels of internal consistency among the subscales of the analytic rubrics for both writing tasks. By contrast, dependability, when examined in terms of information yielded from MG-theory using 'Universe scores' and the dependability coefficient, 'phi', considered an analogue to reliability in CTT (Bachman, 2004; Brennan, 2000), a more complex picture emerged, showing still acceptable, but notably lower estimates of dependability (Grabowski, 2009).

In determining the cause for these lower estimates yielded in MG-theory, one would simply need to refer to the descriptive statistics (i.e., by task and subscale) and the variance components for the current measurement design. That is, the role of *task variability* between the customer review task and the argumentative essay task were clearly present all subscales, but especially so in the case of the 'Content control' scale. The considerable differences in mean rating by task, and the *differential* in rank-ordering of test takers by task, taken together, suggest, not only that the customer review was either less difficult or perceived as less difficult by raters than the argumentative essay, but also that performance on each task did not consistently rank-order test takers from highest to lowest scores (Grabowski & Lin, 2019).

These findings on task-related variability in the writing scores could be considered a trade-off between, on the one hand, sufficiently broad representation of content associated with the construct of L2 writing ability, and, on the other hand, Dependability as conceived of in G-theory (Deville & Chaloub-Deville, 2006; Grabowski, 2009; Schoonen, 2012; Xi & Mollaun, 2006).

Logically, for the former, it is not fair to expect *all* language learner's L2 writing abilities to be consistent across *all* target genres. For instance, they may simply be *more proficient* at writing, say, a customer review, than, say, an argumentative essay, or vice versa. The testing context of the CLP has been reported to serve a wide range of adult ESL students with a variety of linguistic, cultural, educational, and professional backgrounds (Vafaee & Yaghmaeyan, 2020), so it stands to reason that the students that comprise this very test administration may have had differing levels of familiarity, experience, and instruction on how to write an online customer review (e.g., similar to those written on Yelp.com or a similar site) and how to write an argumentative essay (e.g., similar to those written for the TOEFL iBT Independent Writing Task). Considering test-taker variance across such tasks to be 'undesirable' or 'construct-irrelevant' would overlook the fact that *differential* abilities across context are inherent to the process of language learning, differences in cultural exposure among learners, and moreover, familiarity to task format and content, leading some to propose a reconceptualization on task variability in L2 testing to be "ability-in-language user-in-context" (Deville & Chalhoub-Deville, 2006 as cited in Schoonen, 2012, p. 375).

For the latter, the challenge in G-theory accounting for the *construct-relevance* of task variability in test score variance would seem to be a rather fundamental methodological one. That is, if 'tasks' are considered a random facet, and thus interchangeable with others in the *'Universe of Admissible Observations'*, that would suggest that variance attributed to task variability (and not test-taker) would be undesirable because performance across task does not remain consistent (Schoonen, 2012). To address this issue, Xi and Mollaun (2006) suggest the notion of 'cloned' tasks (i.e., similar tasks) (p. 39) as a practical means by which the '*Universe of Admissible Observations'* could be narrowed in order to improve score dependability. However,

while this would address the measurement concern, it would still overlook obvious consideration – that administering the same writing task twice, which would likely provide a more consistent measurement, would narrow the sample test-taker performance to a single written genre. Considering this, there is certainly argument to be made for maintaining a sufficiently broad construct of L2 writing ability at the expense of lower estimates of dependability.

Regardless of this task-related variability, the results indicates that the components of writing ability across task to be highly interrelated and equally contributing to the composite score, which, when considered separately provide basic evidence that they are measuring the same underlying L2 ability, a consideration for making relevant *generalizations* about observed scores on this test (Vafaee & Yaghmaeyan, 2020). Taken together, they could provide a justification for combining scores from different subscales into a composite score to be reported test-takers, which provides evidence to support a particular test score use, discussed next.

With this in mind, the findings from the fifth and sixth research questions are meant to provide empirical backing for the *Utilization Inference*, linking score interpretation to test score use (Chapelle, 2008). Using the aforementioned rationale regarding the highly-related subscales and the nearly identical contributions by task, it seems fair to conclude that the current practice of combining these scores into a composite score (out of 15 points, see Table 5) to be appropriate one.

That said, the obvious caveat that should not be overlooked is that, based on observed performance, a noticeably less difficult (or more leniently rated) customer review task appeared to 'compensate' rather considerably for a much more difficult (or more severely rated) argumentative essay task (Bachman, 2004; Sawaki, 2007). Considering this, it seems that an argument could be made for combining *only* rating within tasks, then creating task-level cut-scores and a profile of composite task scores to report to test takers. Administratively, this adds a layer of complexity. However, it also provides a better representation of writing performance on the placement test.

Lastly, in terms of the findings from the alternative D-studies when varying task and rating conditions, measurements of dependability ( ) changed rather predictably when each was decreased or increased, though more dramatically for tasks. However, decreasing tasks or raters would bring dependability measurements to unacceptably low levels, while increasing them would only cause incremental improvements in dependability.

For this reason, I argue that, in order to balance the needs of test score dependability and practicality (Bachman & Palmer, 1996), it is most appropriate to keep the current measurement design since increasing the number of tasks would add a considerable amount of performance samples to rate (i.e., +208 for this sample) and increasing the number of ratings would entail assigning more work to the current panel or raters, or recruiting more, which seems impractical.

CONCLUSION

Overall, the current study illustrates the use of MG-theory for examining score variability and dependability for written performance assessment on an ESL placement test, rated using an analytic rubric with three subscales. In particular, this study identified the presence of task-related variability that, admittedly, did reduce score dependability for the writing scores yielded from this test, but, by the same token, could substantively be justified as an artifact of representing the construct of L2 writing ability in a sufficiently broad manner. Simply said,

should we really expect test takers to have equivalent levels of proficiency when writing a customer review and an argumentative essay? Moreover, aside from variability associated with the task, the findings provided evidence for maintaining the current measurement design of the writing section, and slightly adjusting the score calculation and reporting procedures for the writing section to account for differences in task difficulty.

ACKNOWLEDGEMENTS

Special thanks to Kirby Grabowski for her support throughout the course, *Generalizability Theory in Second Language Testing*, and to Kimberly Tan and Katherine Gorbenko, who used the same CLP placement test dataset in their term papers, for their willingness to help troubleshoot technical issues and collaboratively make sense of their statistical findings.

REFERENCES

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, U.K.: Cambridge University Press.

Bachman, L.F., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*, 238-257.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Brennan, R. L. (2000). (Mis) Conceptions about generalizability theory. *Educational Measurement, Issues and Practice, 19*(1), 5-10.

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

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Carr, N. (2010). *Designing and analyzing language tests.* Oxford, UK: Oxford University Press.

Chapelle, C. (2008). The TOEFL validity argument. In C. Chapelle M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319-352). New York: Routledge.

Chapelle, C. A. (2011). Validation in language assessment. *The Routledge Handbook of Second Language Acquisition and Language Testing*, 2.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and Generalizability in Applied Linguistics* (pp. 9-25). John Benjamins.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit all? *Language Testing*, *26*(4), 507-531

Grabowski, K. C. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking* (Doctoral dissertation, Teachers College, Columbia University).

Grabowski, K., & Lin, R. (2019). Multivariate generalizability theory in language assessment research. In V. Aryadoust & M. Raquel (Eds.), *Routledge Language Assessment Series*. New York: Routledge.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341-366.
Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th Ed.) (pp. 17-64). Westport, CT: Greenwood Publishing.
Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166.
Lynch, B., & McNamara, T. (1998). Using G theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants.
Purpura, J. E. (2011). Quantitative research methods in assessment and testing. In C. Chapelle & E. Hinkel (Eds.), *Validation in language assessment. Handbook of Research in Second Language Teaching and Learning* (pp. 731-751). Mahwah, NJ: Lawrence Erlbaum Associates.
Sato, T. (2011). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223-241.
Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.
Sawaki, Y., & Sinharay, S. (2013). Investigating the value of section scores for the TOEFL iBT test. *ETS Research Report Series*, *2013*(1), i-113.
Sawaki, Y., & Xi, X. (2019). Univariate generalizability theory in language assessment. *Quantitative data analysis for language assessment*, *1*, 30-53.
Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
Schoonen, R. (2012). The generalisability of scores from language tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 363-377). New York: Routledge.
Vafaee, P., & Yaghmaeyan, B. (2020). Providing evidence for the generalizability of a speaking placement test scores. *International Journal of Language Testing*, *5*(2), 78-95.
Webb, N. M., Shavelson, R. J. & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans (Ed.), *Generalizability theory: Inferences and practical applications* (pp. 67-81). San Francisco, CA: Jossey-Bass.
Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) foroperational use. *Language Testing 24*(2), 251-286.
Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education*, Vol. 7 (pp. 177-196). New York: Springer.
Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series*, *2006*(1), i-71.

Daniel Eskin is a doctoral student in Applied Linguistics at Teachers College, Columbia University. His research articles include the assessment of L2 Pragmatics and Scenario-Based Language Assessment. Correspondence should be sent to E-mail: dae2129@tc.columbia.edu