THE EFFECTS OF PERSONS, ITEMS, AND SUBTESTS ON UH ELIPT READING TEST SCORES

YAO ZHANG

University of Hawai'i

INTRODUCTION

Generalizability Theory

Classical testing approach is typically used to estimate the internal consistency and reliability of a given test, such as K-R 20 and 21, and Cronbach's Alpha. Decision makers also usually look at the related standard error of measurement to estimate the total error of the test. In using these indices, however, there is no way of obtaining a total picture of the combined effects of all relevant sources of error, as pointed out by Erlich and Shavelson (1976, p. 23, cited by Bolus & Hinofotis 1982, p. 248).

As an extension to classical theory, generalizability theory (G-theory), introduced by Cronbach, Rajaratnam, and Gleser (1963), offers a viable solution for estimating test reliability and various sources of errors. According to the G-theory framework provided by Brennen (1983), any particular test is regarded as one sample from a *universe of admissible observations,* which contains *facets*, namely *conditions of measurement*, like an item facet or a subtest facet, and it also contains a *population*, the *objects of measurement*, which can be examinees (usually labeled *persons*), for example. The sources of variance are called *G study variance components.* For example, consider, for instance, a persons by items nested within subtests, or p × (i:s), design. The variance components are associated with the main effects for persons (p), items nested within subtests (i:s), and subtests (s), and the two-way interactions for persons and subtests (ps) and persons by items nested within subtests (pi:s) (after Brennan, 1983). The estimates of those variance components can then be used to investigate various possible measurement designs with various numbers of subtests and items. This can be done using D (Decision) Studies of the relative effects on the generalizability coefficient (analogous to reliability coefficient) of various sources of subtests and items nested within subtests on the test (Brown, 1999, p. 219).

Second Language Studies, 21(2), Spring 2003, pp. 107-128.

The application of G theory in language testing was first suggested in Bolus, Hinofotis, and Bailey (1982) and has been used to investigate oral, written, and reading tests, as well as both criterion- and norm-referenced tests. Bolus et al (1982) looked into the variance components of persons, raters, and occasions in the oral proficiency test of non-native-speaking teaching assistants. Stansfield and Kenyon (1992) investigated the effects of numbers of tests and numbers of raters on the dependability of oral proficiency interview scores. Bachman, Lynch, and Mason (1995) investigated the variability in tasks and rater judgments on a speaking test using G theory and many-faceted Rasch analysis. Brown and Bailey (1984) studied the effects of numbers of raters and numbers of scoring categories on the dependability of writing scores.

G theory has also proven beneficial for the investigation of the dependability of criterion-referenced language tests (see Brown, 1990; 1993; and Kunnan, 1992). The effects of items and subtests were analyzed in Brown (1996) and Brown (1999), which investigated the decision dependability of subtests and the overall TOEFL test battery, as well as the relative importance of persons, items, subtests, and languages to TOEFL test variance. G theory was also applied to investigating the effects of passage/topic variance in Brown's early 1984 work, in which the effects of numbers of items and numbers of passages on the dependability of an engineering English reading test were investigated.

Purpose

The purpose of this study is to investigate the test reliability of the reading test part of the University of Hawaii English Language Institute Placement Test (UH ELI Placement Test) using both classical theory and G-theory. Furthermore, it will investigate the error variance of this reading test as well as how to make the test more efficient and increase its reliability through further rational modifications, such as changing (increasing or decreasing) the number of subtests and/or items. To those ends, this paper will address the following research questions:

1. What do the classical testing analyses show us about the reliability of the subtests and the overall reading test of the ELI Placement Test?
2. What are the G-study variance components for persons, items, and subtests, and their interactions in the test?

3. What are the D-study dependability estimates for various numbers of items and subtests?

METHOD

Participants

A total of 94 participants were involved in this study. They were all students taking the Spring semester 2002 administration of the ELI placement test at UH. The majority of them came from Asian countries such as Japan, Korea, and China (including Hong Kong and Taiwan) and had TOEFL scores between 500 and 600.

Materials

The material used in this study was the reading test part of the University of Hawaii English Language Institute Placement Test (UH ELI Placement Test). This test is administered every semester to newly-admitted undergraduate and graduate students at the University of Hawaii (UH) with TOEFL scores ranging between 500 and 600. The purpose of the UH ELI Placement Test (in combination with information the TOEFL Listening, Grammar, Reading, & Writing subtests) is to decide what level of the listening, reading, and writing courses a student should take at UH, or whether they can be exempted from taking some or all such courses. The test is divided into four parts: listening, grammar, reading, and writing. The reading part consists of a cloze fill-in-the-blank test (CL), and a multiple-choice comprehension test which is composed of a reading comprehension (RC) and an academic vocabulary (AV) test. Due to their different formats (described below) and different constructs, the RC and AV tests will be treated as two separate subtests in the following analyses.

The three subtests in the reading part of the 2002 Spring ELI Placement test battery are as follows:

The *CL Test* is actually a gap-filling test. It is composed of a passage about immigration with 50 blanks selectively, not randomly, deleted. Students are required to write one word in each blank. Three aspects of the issue are discussed in three paragraphs with subtitles given in underlined bold capitalized letters. The answer key was piloted before the test was first carried out and further developed after that. In most cases, there is more than one acceptable answer for each blank. Both the acceptable answers and unacceptable answers are provided for the raters to assist them in manually marking the answer sheets. One point is given for each correct answer. In case there are answers given which are not listed, at least three raters need to agree that it is correct before it is taken to be correct, and the word is then added to the acceptable answer key.

In the *RC Test*, there are six passages of 200 to 350 words each written in academic style with four or five comprehension questions following each one. There are four options for each item and students must circle the letter for the best answer on the answer sheet. One point will be given to each correctly answered item with no penalty for guessing.

The *AV Test* is designed on the assumption that the knowledge of academic vocabulary is one important component of the academic reading ability. In the prompt, either a word or a definition is given. Four options are provided for each item to choose from. Options can be definitions, explanations, synonyms, or words/phrases. The task is to choose one of the four options that best match the meaning of the words/definitions given in each prompt. One point is given for each correctly answered question, and again, there is no penalty for guessing.

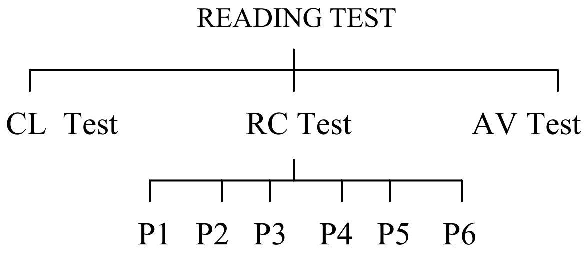
Procedures

The test took place in the morning from 8:00 am to 12:00 am with a break in between. The first writing test was only taken by graduate students, and the rest by both graduate and undergraduate students. The reading test took up the second half of the test with 35 minutes for the CL test, and 50 minutes for the RC and AV tests combined. Strict admission procedures were followed. The answer sheets were collected and scored immediately after the test and students could get their results on a website the next morning.

Analyses

The analyses in this study started with descriptive statistics and classical theory reliability estimates (specifically Cronbach alphas) used to compare with the results of the generalizability studies (G-studies) and decision studies (D-studies) that followed. Five G-studies were conducted using the computer program GENOVA (Crick & Brennan, 1982, cited by Brennan 1983, p. 141). The overall structure of these five G-studies is illustrated in Figure 1.

Figure 1: Overall Structure of Five G-studies



The first and second G-studies considered the effect of numbers of items and subtests on the dependability of the total Reading Test battery scores and of the RC subtest scores, respectively. The subtests for the Reading Test are the CL, RC, and AV subtests and those for the RC subtest are passages One to Six (P1 to P6). The third, fourth, and fifth G-studies investigated the effect of numbers of items on the CL, RC, and AV tests, respectively. A two-facet design (an item facet and a subtest facet) was used to carry out the first and second G-studies, which is a persons by items nested within subjects design, or p × (i:s). The third to fifth G-studies adopted a single-facet (item facet) design, which was a persons crossed with items design, or p × i. Mixed models were used in the first and second G-studies in order to examine the current configuration of the reading test with fixed effects for the subtest facet and random effects for the item facet. Random effects models were used in all the five G-studies to investigate how the results would generalize to other measurement conditions, in that the current persons, items, subtests, and/or passages were regarded as samples from a universe of admissible observations.

Since unbalanced G-studies (as studies with different numbers of items in each subtest are called) make estimation of the variance components computationally very complex (Brown 1999, p. 222), the first 25 items were selected from the CL Test so that it would have the same number of items as the RC Test and AV Test in the first G-study. Similarly, the first four items (out of five) were chosen from the fourth passage in the RC Test so that it would have the same number of items as the other passages.

Variance components were then estimated using the GENOVA computer program in order to understand the relative contributions of persons, items, and subtests to the error variance.

Next, D-studies were conducted parallel to these five G-studies, which emphasized the estimation, use, and interpretation of variance components for decision-making with well-specified measurement procedures (Brennan, 1983, p. 3). Only "relative" error variance $\sigma^2(\delta)$ was considered because the tests involved here are all designed for norm-referenced decisions.

Corresponding generalizibility coefficients (based on $\sigma^2(\delta)$) are reported for various numbers of items and subtests in the first two D-studies (parallel to the first two G-studies) so that the reader can directly observe the effect on dependability of those facets in various combinations (Brown, 1999, p. 237).

RESULTS

Descriptive Statistics and Classical Theory Reliability

To answer the first research question, the raw score descriptive statistics and reliability estimates are given in Table 1 for the original data as well as the sub-samples used for G-studies. For the first G-study, 75 items were gathered, consisting of the first 25 items of the CL Test and all the items from the RC Test and AV Test (25 items each). For the second G-study, 24 items were selected from the RC Test, four items from each of the six passages. The fifth item of the fourth passage was dropped for the sake of having a balanced design for the G-study.

Generally speaking, the Cronbach Alpha (α) reliability estimates for the overall original reading test and the G-study sampling in the first G-study are acceptably high

(.91 and .87). The difference is probably due to the reduction in numbers of items. As for the CL Test, when the number of items was reduced by half, the reliability dropped from .85 to .71, which might be due to the change in the number of items. An interesting finding in the RC Test was that the reliability increased from .67 to .69 when the fifth item of P4 was dropped. That item was apparently a weak item which reduced the reliability of the test. In the G-study sampling, the AV Test had the highest reliability estimate (.81) among the three subtests. When looking at the standard deviation and the range of the three subtests, we see that the AV Test scores have the highest standard deviation (5.16) and widest range (2-25), the RC Test has the lowest standard deviation (3.91) and narrowest range (5-22), and the CL Test is in the middle (4.10 for standard deviation and 2-22 for the range). This indicates that the AV Test distributed persons comparatively well, but the RC Test did poorly in this respect.

Five G-studies (Persons by Items Nested Within Subtests, or p × (i:s) Designs, and Persons Across Items, or p × i Designs)

The first two G-studies were identical in structure. They were for persons by items nested within subtests, or p × (i:s) design. Such a design reveals the relative contributions of persons, items, and subtests, and their interactions in terms of variance components. The remaining three G-studies are persons crossed with items, or p × i, designs, in which the relative variance components for persons and items were examined. All the analyses were conducted using GENOVA, a computer program specially designed for analyzing G-studies.

Table 1

Descriptive Statistics and Classical Theory Reliability Estimates

(for the original Reading test set and the corresponding G-study sampling)

STUDY **TEST**	**ORIGINAL TEST**					**G-STUDY SAMPLING**				
SUBTEST **PASSAGE**	MEAN	*SD*	RANGE	ALPHA	*k*	MEAN	*SD*	RANGE	ALPHA	*K*
STUDY ONE										
READING	5.45	14.90	16-87	.91	100	4.46	1.94	16-67	.87	75
CL	22.51	8.11	2-45	.85	50	12.52	4.10	2-22	.71	25
RC	14.14	3.91	5-22	.67	25	14.14	3.91	5-22	.67	25
AV	13.80	5.16	2-25	.81	25	13.80	5.16	2-25	.81	25
STUDY TWO										
RC	14.14	3.89	5-22	.67	25	14.05	3.91	5-22	.69	24
P1	1.84	1.00	0-4	.15	4	1.84	1.00	0-4	.15	4
P2	2.44	1.00	0-4	.17	4	2.44	1.00	0-4	.17	4
P3	2.50	1.05	0-4	.39	4	2.50	1.05	0-4	.39	4
P4	3.06	1.07	0-5	.34	5	2.98	1.11	0-4	.52	4
P5	2.24	1.01	0-4	.04	4	2.24	1.01	0-4	.04	4
P6	2.05	1.16	0-4	.37	4	2.05	1.16	0-4	.37	4
STUDY THREE										
CL	22.51	8.11	2-45	.85	50	22.51	8.11	2-45	.85	50
STUDY FOUR										
RC	14.14	3.91	5-22	.67	25	14.14	3.91	5-22	.67	25
STUDY FIVE										
AV	13.80	5.16	2-25	.81	25	13.80	5.16	2-25	.81	25

Table 2

Variance Components for the First Two G-studies

Source	Study One Reading Test		Study Two: RC Test	
	Raw Components	Percentage	Raw Components	Percentage
σ^2 (p)	.0162624	6.53%	.0181776	7.43%
σ^2 (s)	(0.0)[a]	0.00%	.0045944	1.88%
σ^2 (i:s)	.0251952	10.12%	.0185821	7.60%
σ^2 (ps)	.0070751	2.84%	.0020250	0.83%
σ^2 (pi:s)	.2004785	80.50%	.2011283	82.26%
Total	.2490112	99.99%[b]	.2445074	100.00%

a. This value was a negative variance component, which was rounded to zero after Brennan (1983)

b. The total percent is not 100 because of the rounding

In Table 2, we can see that the variance components for the Reading Test and the RC Test have similar trends. The percentage of variance accounted for by persons in the Reading Test is reasonably high (σ^2_p=6.53%), but it is lower than the percentage accounted for by items variance ($\sigma^2_{i:s}$=10.12%), which means that people performed differently across items in the Reading Test, but only to some degree. In fact, the variance component for items is the single largest main effect ($\sigma^2_{i:s}$=10.12%), which probably reveals that the items had a wide range of difficulties. In contrast, the percentage of variance due to subtests on the Reading Test is negligible (σ^2_s=0.00%), which means all three subtests were about the same in difficulty. The two-way interaction between persons and subtests (*ps*) is relatively small ($\sigma^2_{ps=}$2.84%). So we can conclude that persons performances varied only to a small degree across the three subtests. The lion's share of variance is found in the two-way interaction between persons and items nested within subtests ($\sigma^2_{pi:s=}$80.50%). This shows that persons performed very differently for different items. In another word, different persons answered different items correctly.

Similar to the distribution of variance components on the Reading Test, on the RC subtest, persons variance was relatively low (σ^2_p=7.43%). Difficulty across the six passages differed slightly (σ^2_s=1.88%) and items behaved differently in the sense of difficulty ($\sigma^2_{i:s}$=7.6%). To a very small degree (σ^2_{ps}=0.83%), persons performed differently across passages, but persons again performed very differently on different items ($\sigma^2_{pi:s}$=82.26%). The finding that items on the overall Reading Test vary in difficulty demands further investigation into what the variance components for persons and items are for each subtest.

Table 3

Variance Components for the Third to Fifth G-Studies

Source	Study Three The CL Test		Study Four The RC Test		Study Five The AV Test	
	Raw Components	Percentage	Raw Components	Percentage	Raw Components	Percentage
σ^2 (p)	.022324	8.99%	.016521	6.68%	.034468	13.89%
σ^2 (i)	.027075	10.90%	.031605	12.78%	.009049	3.65%
σ^2 (pi)	.198943	80.11%	.199105	80.53%	.204603	82.46%
Total	.248343	100.00%	.24723	99.99%*	.248121	100.00%

*The total percent is not 100 because of the rounding

In Table 3, examining the three subtests separately, the interactions between persons and items contribute the largest variance components (80.11%, 80.53%, and 82.46% for the CL Test, the RC Test, and the AV Test, respectively). This means, in general, different persons performed differently on different items. Such differences are most obvious in the AV Test (σ^2_{pi}=82.46%) and a little less in the CL Test (σ^2_{pi}=80.11%). Great divergence appears in terms of persons variance and items variance across three subtests. It turned out that the percentage of the persons variance was the greatest in the AV Test (σ^2_p=13.89%), and smallest in the RC Test (σ^2_p=6.68%), with the CL test in the middle (σ^2_p=8.99%). This means that the AV Test best distributed persons (desirable for a norm-referenced test) and the RC Test did relatively poorly in spreading people out. The CL Test behaved in the middle when judged by the dispersion of persons. Judged from the percentage of the item variance components, the difficulty of items varied least in the AV Test (σ^2_i=3.65%), almost three times as much in the CL Test (σ^2_i=1.90%), and varied most in the RC Test (σ^2_i=12.78%). When the three subtests are judged individually, the AV Test appears to be the most effective subtest in terms of reliability in the sense that it most widely distributed persons across the items with similar difficulties. The RC Test is the poorest in this respect and the CL Test is in the middle.

D-study Results and Generalizability Coefficients

Table 4 presents the results of the D-studies parallel to the five G-studies, first on a random effects model for all five studies and then on a mixed effects model for the first two studies. The random effects model estimates allow generalization of the results to other tests. The statistics for this model include $\sigma^2(\tau)$, which is another expression of $\sigma^2(p)$, and the lower-case delta error term, $\sigma^2(\delta)$, (for relative decisions, i.e., norm-referenced interpretations). Also, the G-coefficients, $E\rho^2(\delta)$, analogous to reliability coefficients, are presented in Table 4. They were calculated by forming the ratio of the persons variance component for the particular number of subsets and items in the G-study over the same persons variance plus the appropriate error term (Brown, 1996, p. 244). Thus, G-coefficients for relative decisions would use δ error as follows:

$$E\rho^2(\delta)= \frac{\sigma^2(\tau)}{\sigma^2(\tau)+ \sigma^2(\delta)} = \frac{\sigma^2(p)}{\sigma^2(p)+ \sigma^2(\delta)}$$

In the second part of the table, the same statistics are presented for a mixed effects model (with subtests as a fixed effect). These results can only be generalized to the Reading test battery as it was structured and studied here.

Table 4

D-study results for

STUDY **MODEL STATISTICS**	**STUDY ONE READING TEST**	**STUDY TWO RC TEST**	**STUDY THREE CL TEST**	**STUDY FOUR RC TEST**	**STUDY FIVE AV TEST**
DESIGN	p × (i:s) n_s=3 n_i=25 $n_i n_s$ =75	p × (i:s) n_s=6 n_i=4 $n_i n_s$ =24	p × i n_i=50	p × i n_i=25	p × i n_i=25
RANDOM EFFECTS MODEL					
σ^2 (p)	.0163	.0182	.0223	.0165	.0345
$\sigma^2(\tau)$	.0163	.0182	.0223	.0165	.0345
$\sigma^2(\delta)$	.0050	.0087	.0040	.0080	.0082
$E\rho^2(\delta)$	.7637	.6759	.8487	.6747	.8081
MIXED EFFECTS MODEL					
$\sigma^2(\tau)$	.0186	.01852			
$\sigma^2(\delta)$	.0027	.0084			
$E\rho^2(\delta)$	.8745	.6884			

Table 5

Generalizability Coefficients for the Reading Test

I/S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	30	35	40	45	50	60	75	100
1	.073	.135	.190	.239	.281	.320	.354	.385	.414	.439	.463	.485	.505	.523	.540	.556	.571	.585	.598	.610	.622	.633	.643	.653	.662	.702	.733	.758	.779	.797	.825	.855	.887
2	.132	.233	.313	.377	.431	.476	.515	.548	.577	.602	.625	.645	.663	.680	.694	.708	.720	.732	.742	.752	.761	.769	.777	.784	.791	.820	.841	.858	.872	.883	.901	.919	.938
3	.180	.306	.398	.468	.524	.569	.606	.638	.664	.688	.708	.725	.741	.755	.767	.779	.789	.798	.807	.815	.822	.829	.835	.841	.846	.868	.885	.898	.908	.917	.930	.943	.957
4	.221	.363	.460	.532	.587	.630	.666	.695	.719	.740	.758	.773	.787	.799	.810	.820	.829	.837	.844	.850	.857	.862	.867	.872	.877	.895	.909	.919	.928	.934	.945	.955	.966
5	.256	.408	.508	.580	.633	.674	.707	.734	.756	.775	.791	.805	.818	.828	.838	.847	.854	.861	.868	.873	.879	.884	.888	.892	.896	.912	.923	.932	.939	.945	.954	.963	.972
6	.287	.445	.546	.616	.668	.707	.738	.763	.783	.801	.815	.828	.839	.849	.858	.865	.872	.878	.884	.889	.894	.898	.902	.906	.909	.923	.934	.941	.948	.953	.960	.968	.976
7	.313	.477	.577	.646	.695	.732	.761	.785	.804	.820	.834	.845	.855	.864	.872	.879	.886	.891	.896	.901	.905	.909	.913	.916	.919	.932	.941	.948	.953	.958	.965	.972	.979
8	.336	.503	.603	.669	.717	.752	.780	.802	.820	.835	.848	.859	.868	.876	.884	.890	.896	.901	.906	.910	.914	.918	.921	.924	.927	.938	.947	.953	.958	.962	.968	.974	.981
9	.357	.526	.624	.689	.735	.769	.795	.816	.833	.847	.859	.869	.878	.886	.893	.899	.904	.909	.913	.917	.921	.924	.927	.930	.933	.943	.951	.957	.961	.965	.971	.977	.982
10	.375	.545	.643	.706	.750	.782	.808	.827	.844	.857	.868	.878	.886	.894	.900	.906	.911	.915	.919	.923	.926	.930	.932	.935	.937	.947	.955	.960	.964	.968	.973	.978	.984
11	.391	.562	.659	.720	.763	.794	.818	.837	.853	.865	.876	.885	.893	.900	.906	.911	.916	.920	.924	.928	.931	.934	.937	.939	.941	.951	.957	.963	.967	.970	.975	.980	.985
12	.406	.578	.672	.732	.774	.804	.827	.845	.860	.872	.883	.891	.899	.905	.911	.916	.921	.925	.929	.932	.935	.938	.940	.943	.945	.954	.960	.965	.969	.972	.976	.981	.986
13	.420	.591	.684	.743	.783	.813	.835	.853	.867	.878	.888	.897	.904	.910	.916	.920	.925	.929	.932	.935	.938	.941	.943	.946	.948	.956	.962	.967	.970	.973	.977	.982	.986
14	.432	.603	.695	.753	.792	.820	.842	.859	.872	.884	.893	.901	.908	.914	.919	.924	.928	.932	.935	.938	.941	.944	.946	.948	.950	.958	.964	.968	.972	.974	.979	.983	.987
15	.443	.614	.705	.761	.799	.827	.848	.864	.877	.888	.897	.905	.912	.918	.923	.927	.931	.935	.938	.941	.944	.946	.948	.950	.952	.960	.965	.970	.973	.975	.979	.984	.988
16	.453	.624	.713	.768	.806	.833	.853	.869	.882	.892	.901	.909	.915	.921	.926	.930	.934	.937	.940	.943	.946	.948	.950	.952	.954	.961	.967	.971	.974	.976	.980	.984	.988
17	.463	.633	.721	.775	.812	.838	.858	.873	.886	.896	.905	.912	.918	.923	.928	.932	.936	.939	.942	.945	.948	.950	.952	.954	.956	.963	.968	.972	.975	.977	.981	.985	.989
18	.472	.641	.728	.781	.817	.843	.862	.877	.889	.899	.908	.915	.921	.926	.931	.935	.938	.941	.944	.947	.949	.952	.954	.955	.957	.964	.969	.973	.976	.978	.982	.985	.989
19	.480	.649	.735	.787	.822	.847	.866	.881	.893	.902	.910	.917	.923	.928	.933	.937	.940	.943	.946	.949	.951	.953	.955	.957	.958	.965	.970	.974	.976	.979	.982	.986	.989
20	.487	.655	.740	.792	.826	.851	.869	.884	.895	.905	.913	.919	.925	.930	.935	.938	.942	.945	.948	.950	.952	.954	.956	.958	.960	.966	.971	.974	.977	.979	.983	.986	.990
21	.495	.662	.746	.796	.830	.854	.873	.887	.898	.907	.915	.922	.927	.932	.936	.940	.943	.946	.949	.951	.954	.956	.957	.959	.961	.967	.972	.975	.978	.980	.983	.987	.990
22	.501	.668	.751	.801	.834	.858	.876	.889	.900	.909	.917	.923	.929	.934	.938	.941	.945	.948	.950	.953	.955	.957	.959	.960	.962	.968	.972	.976	.978	.980	.984	.987	.990
23	.507	.673	.755	.805	.837	.861	.878	.892	.903	.911	.919	.925	.931	.935	.939	.943	.946	.949	.951	.954	.956	.958	.959	.961	.963	.969	.973	.976	.979	.981	.984	.987	.990
24	.513	.678	.760	.808	.841	.863	.881	.894	.905	.913	.921	.927	.932	.937	.941	.944	.947	.950	.952	.955	.957	.959	.960	.962	.963	.969	.974	.977	.979	.981	.984	.988	.991
25	.519	.683	.764	.812	.843	.866	.883	.896	.907	.915	.922	.928	.933	.938	.942	.945	.948	.951	.953	.956	.958	.960	.961	.963	.964	.970	.974	.977	.980	.982	.985	.988	.991
30	.542	.703	.780	.825	.855	.876	.892	.904	.914	.922	.929	.934	.939	.943	.947	.950	.953	.955	.957	.959	.961	.963	.965	.966	.967	.973	.976	.979	.982	.983	.986	.989	.992
35	.560	.718	.792	.836	.864	.884	.899	.910	.920	.927	.933	.938	.943	.947	.950	.953	.956	.958	.960	.962	.964	.965	.967	.968	.969	.974	.978	.981	.983	.985	.987	.990	.992
40	.574	.729	.801	.843	.871	.890	.904	.915	.924	.931	.937	.942	.946	.950	.953	.956	.958	.960	.962	.964	.966	.967	.969	.970	.971	.976	.979	.982	.984	.985	.988	.990	.993
45	.585	.738	.809	.849	.876	.894	.908	.919	.927	.934	.939	.944	.948	.952	.955	.958	.960	.962	.964	.966	.967	.969	.970	.971	.972	.977	.980	.983	.984	.986	.988	.991	.993
50	.595	.746	.815	.854	.880	.898	.911	.921	.930	.936	.942	.946	.950	.954	.957	.959	.961	.964	.965	.967	.969	.970	.971	.972	.973	.978	.981	.983	.985	.987	.989	.991	.993
60	.610	.757	.824	.862	.886	.904	.916	.926	.934	.940	.945	.949	.953	.956	.959	.962	.964	.966	.967	.969	.970	.972	.973	.974	.975	.979	.982	.984	.986	.987	.989	.992	.994
75	.625	.769	.833	.870	.893	.909	.921	.930	.938	.943	.948	.952	.956	.959	.962	.964	.966	.968	.969	.971	.972	.973	.975	.976	.977	.980	.983	.985	.987	.988	.990	.992	.994
100	.642	.782	.843	.878	.900	.915	.926	.935	.942	.947	.952	.956	.959	.962	.964	.966	.968	.970	.971	.973	.974	.975	.976	.977	.978	.982	.984	.986	.988	.989	.991	.993	.994

As would be expected, the G-coefficients E$\rho^2(\delta)$ for the mixed model of the Reading and RC Tests are very similar to the Cronbach Alpha values reported in Table 1 for the G-study sampling (.87 and .69, respectively), but slightly different from the corresponding Cronbach Alpha for the original test (.91 and .67, respectively). This may be due to differences in the number of items.

Naturally, the G-coefficients in the random model are much more conservative than those for the mixed model. This is because the random effects statistics can be generalized beyond the items and subtests of the current Reading Test to other batteries and tests. Tables 5 and 6 present the random effects G-coefficients that would arise from different numbers of items and subtests. The top row indicates the numbers of subtests and the left column shows the numbers of items.

For example, Table 5 for the total Reading Test battery shows that the G-coefficient for three subtests with 25 items each is .764 (at the point where the 25th row and the third column of coefficients intersect), which is equivalent to the random effects model G-coefficient of .7637 reported in Table 4. Notice that the G-coefficient would be .625 if the battery were configured with the same 75 items but with one subtest. With five subtests of 15 items each, it is predicted that it would be dependable at .799, and with 15 subtests of five items each, it would be .838. Thus, the effects of having the items divided up into smaller and smaller subtests are clearly illustrated.

As is seen, there is a considerable gain in dependability from having the Reading test made up of three different subtests instead of one long homogeneous test. In other words, dependability increases due to increases in the number of subtests involved, while holding the number of items constant. Such increases are above and beyond predictions that could be made by classical theory reliability studies.

Table 5 also provides information for other combinations of the number of items and subtests, which will be helpful in any future revisions of this or other tests. For instance, in a test with five subtests and 12 items each, the dependability would be .774, which is higher than the current test configuration. Moreover, the total number of items would then be 60, which is 15 items less than the G-study sample. In other words, the dependability would increase with a smaller number of items. This reveals the flexibility available for modifying a test based on the results of D-studies.

Nevertheless, practicality should be taken into consideration as far as actual decisions to modify the tests are concerned. For example, a test with 100 subtests of 11 items each is predicted to be dependable at .99, but such an 1100-item test is far from practical even though the dependability would be nearly perfect. As Brown (1996, p. 246) pointed out, these dependability estimates for various numbers of items and subtests are meant to provide one piece of information among the many types of information that must be considered in making test development decisions.

Referring to Table 6 for the RC Test, a single test with 24 items would be dependable at .636 while a similar 24 item test based on two subtests of 12 items each would only be slightly more dependable at .659. Three subtests with eight items each would only gain .004 points at .68. In short, the pay off in terms of subtests (while the number of items is held constant) seems to be minimal for the RC Test.

Recall that the item facet is a very important source of variance in the Reading Test and the RC Test. Therefore, the researcher feels obligated to investigate the impact on the dependability of the changing number of items in each of the subtests. The results show different pictures for the two tests discussed here. Based on the current test configuration, for example, in the Reading Test, suppose five items are added to each of the subtests. The dependability would increase from .764 to .780, a gain of .016. For the RC Test, even adding one item to each of the six passages would cause an increase in the dependability from .676 to .721, a gain of .045. This shows that the RC Test will benefit more than the total Reading Test from increasing the number of items.

Table 6

Generalizability Coefficients for the RC Test

I/S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	30	35	40	45	50	60	75	100
1	.082	.152	.212	.264	.309	.349	.385	.417	.446	.472	.496	.518	.538	.556	.573	.589	.603	.617	.630	.642	.653	.663	.673	.682	.691	.729	.758	.782	.801	.817	.843	.870	.899
2	.151	.262	.347	.415	.470	.515	.554	.586	.615	.639	.661	.680	.697	.713	.727	.739	.751	.761	.771	.780	.788	.796	.803	.810	.816	.842	.861	.876	.889	.899	.914	.930	.947
3	.208	.345	.441	.513	.568	.612	.648	.678	.703	.725	.743	.760	.774	.787	.798	.808	.817	.826	.833	.840	.847	.853	.858	.863	.868	.888	.902	.913	.922	.929	.940	.952	.963
4	.258	.410	.510	.582	.635	.676	.709	.735	.758	.777	.793	.807	.819	.830	.839	.848	.855	.862	.868	.874	.879	.884	.889	.893	.897	.912	.924	.933	.940	.946	.954	.963	.972
5	.301	.463	.563	.632	.683	.721	.751	.775	.795	.811	.826	.838	.848	.858	.866	.873	.880	.886	.891	.896	.900	.904	.908	.912	.915	.928	.938	.945	.951	.956	.963	.970	.977
6	.338	.506	.605	.672	.719	.754	.782	.804	.822	.836	.849	.860	.869	.877	.885	.891	.897	.902	.907	.911	.915	.918	.922	.925	.927	.939	.947	.953	.958	.962	.968	.975	.981
7	.371	.542	.639	.703	.747	.780	.805	.825	.842	.855	.867	.876	.885	.892	.899	.904	.909	.914	.918	.922	.925	.929	.931	.934	.937	.947	.954	.959	.964	.967	.973	.978	.983
8	.401	.572	.667	.728	.770	.801	.824	.843	.858	.870	.880	.889	.897	.904	.909	.915	.919	.923	.927	.930	.934	.936	.939	.941	.944	.953	.959	.964	.968	.971	.976	.980	.985
9	.427	.599	.691	.749	.789	.817	.839	.856	.870	.882	.891	.900	.907	.913	.918	.923	.927	.931	.934	.937	.940	.943	.945	.947	.949	.957	.963	.968	.971	.974	.978	.982	.987
10	.451	.622	.711	.767	.804	.831	.852	.868	.881	.891	.900	.908	.914	.920	.925	.929	.933	.937	.940	.943	.945	.948	.950	.952	.954	.961	.966	.970	.974	.976	.980	.984	.988
11	.472	.642	.729	.782	.817	.843	.862	.877	.890	.900	.908	.915	.921	.926	.931	.935	.938	.942	.944	.947	.949	.952	.954	.956	.957	.964	.969	.973	.976	.978	.982	.985	.989
12	.492	.659	.744	.795	.829	.853	.871	.886	.897	.906	.914	.921	.926	.931	.936	.939	.943	.946	.948	.951	.953	.955	.957	.959	.960	.967	.971	.975	.978	.980	.983	.986	.990
13	.510	.675	.757	.806	.839	.862	.879	.893	.903	.912	.920	.926	.931	.936	.940	.943	.946	.949	.952	.954	.956	.958	.960	.961	.963	.969	.973	.977	.979	.981	.984	.987	.990
14	.526	.689	.769	.816	.847	.869	.886	.899	.909	.917	.924	.930	.935	.939	.943	.947	.950	.952	.955	.957	.959	.961	.962	.964	.965	.971	.975	.978	.980	.982	.985	.988	.991
15	.541	.702	.779	.825	.855	.876	.892	.904	.914	.922	.928	.934	.939	.943	.946	.950	.952	.955	.957	.959	.961	.963	.964	.966	.967	.972	.976	.979	.981	.983	.986	.989	.992
16	.555	.714	.789	.833	.862	.882	.897	.909	.918	.926	.932	.937	.942	.946	.949	.952	.955	.957	.959	.961	.963	.965	.966	.968	.969	.974	.978	.980	.982	.984	.987	.989	.992
17	.567	.724	.797	.840	.868	.887	.902	.913	.922	.929	.935	.940	.945	.948	.952	.955	.957	.959	.961	.963	.965	.967	.968	.969	.970	.975	.979	.981	.983	.985	.987	.990	.992
18	.579	.734	.805	.846	.873	.892	.906	.917	.925	.932	.938	.943	.947	.951	.954	.957	.959	.961	.963	.965	.967	.968	.969	.971	.972	.976	.980	.982	.984	.986	.988	.990	.993
19	.590	.742	.812	.852	.878	.896	.910	.920	.928	.935	.941	.945	.949	.953	.956	.958	.961	.963	.965	.966	.968	.969	.971	.972	.973	.977	.981	.983	.985	.986	.989	.991	.993
20	.601	.751	.819	.858	.883	.900	.913	.923	.931	.938	.943	.948	.951	.955	.958	.960	.962	.964	.966	.968	.969	.971	.972	.973	.974	.978	.981	.984	.985	.987	.989	.991	.993
21	.610	.758	.825	.862	.887	.904	.916	.926	.934	.940	.945	.950	.953	.956	.959	.962	.964	.966	.968	.969	.971	.972	.973	.974	.975	.979	.982	.984	.986	.987	.989	.992	.994
22	.619	.765	.830	.867	.891	.907	.919	.929	.936	.942	.947	.951	.955	.958	.961	.963	.965	.967	.969	.970	.972	.973	.974	.975	.976	.980	.983	.985	.987	.988	.990	.992	.994
23	.628	.771	.835	.871	.894	.910	.922	.931	.938	.944	.949	.953	.956	.959	.962	.964	.966	.968	.970	.971	.973	.974	.975	.976	.977	.981	.983	.985	.987	.988	.990	.992	.994
24	.636	.777	.840	.875	.897	.913	.924	.933	.940	.946	.951	.954	.958	.961	.963	.965	.967	.969	.971	.972	.973	.975	.976	.977	.978	.981	.984	.986	.987	.989	.991	.992	.994
25	.644	.783	.844	.878	.900	.915	.927	.935	.942	.948	.952	.956	.959	.962	.964	.967	.968	.970	.972	.973	.974	.975	.976	.977	.978	.982	.984	.986	.988	.989	.991	.993	.994
30	.676	.806	.862	.893	.912	.926	.936	.943	.949	.954	.958	.962	.964	.967	.969	.971	.973	.974	.975	.977	.978	.979	.980	.980	.981	.984	.986	.988	.989	.990	.992	.994	.995
35	.701	.824	.875	.903	.921	.933	.942	.949	.955	.959	.963	.966	.968	.970	.972	.974	.975	.977	.978	.979	.980	.981	.982	.983	.983	.986	.988	.989	.991	.992	.993	.994	.996
40	.720	.838	.885	.912	.928	.939	.947	.954	.959	.963	.966	.969	.971	.973	.975	.976	.978	.979	.980	.981	.982	.983	.983	.984	.985	.987	.989	.990	.991	.992	.994	.995	.996
45	.737	.848	.894	.918	.933	.944	.951	.957	.962	.966	.969	.971	.973	.975	.977	.978	.979	.981	.982	.982	.983	.984	.985	.985	.986	.988	.990	.991	.992	.993	.994	.995	.996
50	.750	.857	.900	.923	.938	.947	.955	.960	.964	.968	.971	.973	.975	.977	.978	.980	.981	.982	.983	.984	.984	.985	.986	.986	.987	.989	.991	.992	.993	.993	.994	.996	.997
60	.772	.871	.910	.931	.944	.953	.959	.964	.968	.971	.974	.976	.978	.979	.981	.982	.983	.984	.985	.985	.986	.987	.987	.988	.988	.990	.992	.993	.993	.994	.995	.996	.997
75	.794	.885	.921	.939	.951	.959	.964	.969	.972	.975	.977	.979	.980	.982	.983	.984	.985	.986	.987	.987	.988	.988	.989	.989	.990	.991	.993	.994	.994	.995	.996	.997	.997
100	.818	.900	.931	.947	.957	.964	.969	.973	.976	.978	.980	.982	.983	.984	.985	.986	.987	.988	.988	.989	.990	.990	.990	.991	.991	.993	.994	.994	.995	.996	.996	.997	.998

DISCUSSION

In interpreting the above results, it is important to keep in mind that the dependability estimates of the first G-study were based on 25 fewer items than the original test for a balanced G-study with equal numbers of items on each subtest. Since shorter tests tend to be less reliable, the dependability estimates would be conservative underestimates and not be an overestimate of the real test reliability.

The remainder of the discussion will be presented in the same order as the research questions, which will be used as headings.

What Do the Classical Testing Analyses Show Us about the Reliability of the Subtests and the Overall Reading Test of ELI Placement Test?

The results from Table 1 indicate that the overall Reading Test has high reliability at .91, based on Cronbach alpha. The CL Test and the AV Tests are less reliable at .85 and .81, respectively, which is probably, at least in part, because they are shorter. The RC Test has the least reliability at .67. This might be due to the effect of unreliable item(s) and/or passage(s) (like item 16), which, when taken out of the test, caused an increase in the reliability of Passage 4, as well as the whole RC Test. Moreover, Cronbach alpha for Passage 4 is only .04, which is almost no reliable at all. Notice the standard deviation and the range of scores in the RC Test (3.91 and 5-22, respectively) indicate the distribution is tight and narrow compared to other subtests. Table 2 also presented the same classical theory statistics for the items used in the G-study sampling (done for balanced designs). The Cronbach alpha estimates later turned out to be comparable to the G-coefficients (for δ error) for the mixed models, as would be expected. This result is parallel to what Brown (1996) found.

What Are the Variance Components of Persons, Items, Subtests, and Their Interactions in The Test?

Examining the variance components for five G-studies shown in Tables 2 and 3 reveals the relative contributions to error of persons, subtests (only in Table 2), items (nested within the subtests in Table 2), and their interactions. Generally speaking, the items variance component was the single most important main effect, except in the AV Test. Persons variance in the first four G-studies was considerably higher, though it did not exceed items variance. In the AV Test on the other hand, persons variance was the largest single main effect and item variance was about one-third less than persons variance, which is what it should be in a norm-referenced test (Brown, 1996, 1999). The largest variance components in all of these studies are the interactions between persons and items (nested within subtests in first two G-studies). This indicates that there were considerable differences in persons performance across items, and the reason for this could be that the test items had varying degrees of difficulty, and persons had different language proficiency levels on different items relative to each other. The relative magnitudes of the variance components for the subtests main effect and the interaction of persons and subtests were only moderately high, indicating that subtests and their interactions with persons were not very important relative to all the other sources of variance in this design (i.e., persons, items, and persons by items). Thus, combined together, the findings of p × (i:s) design for the first and second G-studies, and p × i for the third to fifth studies indicate that the examinees' relative proficiency differed considerably across items, but not so much across subtests.

What is the D-Study Dependability for Various Numbers of Items and Subtests?

Tables 5 and 6 for first and second G-studies provide direct answers to this research question. The subtest facet, across all cases, clearly had some effect on the predicted dependability indices since in no D-study was the dependability the same for one subtest and more than one subtest with the number of items held constant. The fact is that dependability was gained by increasing the number of subtests even though the number of items was kept constant.

However, the degree of gain in the dependability achieved by having more subtests with the same total number of items is different between two studies. The influence of subtests was large in Study One (the Reading Test), but relatively small in Study Two (the RC Test).

The variation of items also appears to have an important influence on the dependability, which is also predicted by the classical testing theory. That is, more items tend to enhance reliability and fewer items tend to decrease the reliability of a test. Nevertheless, the degree of gain in dependability achieved by increasing the number of items differs in the D-studies. The RC Test increases in dependability to a greater degree than the Reading Test when increasing the number of items on each subtest.

Also, it should be noted that the two studies varied from each other in structure. The relatively large differences in dependability due to the subtest facet in Study One were due to differences between the tests (i.e., the CL Test, RC Test, and AV Test), while those observed for Study Two were due to differences between reading passages (i.e., P1 to P6).

CONCLUSION

The five G-studies and D-studies conducted here reveal a broad picture of the relative contributions of persons, items, and subtests to the error variance of the Reading Test. The Reading Test is problematic to some degree in that the variance due to items is larger than that for persons, which is also true for the CL test and the RC Test. In the AV Test, however, the persons variance component is the largest single main effect, almost four times as large as the items variance component. This explains why the AV Test has higher reliability than the other two subtests in the G-sampling. Closer examination and refinement should be performed on individual items in the CL test and the RC test, since the difficulty of the items in these two subtests turned out to vary considerably.

The variance due to subtests is only a very small part of the overall variance components indicating that the subtests involved may be testing very much the same thing. This would be a supporting argument for making decisions about students' reading ability based on the complementary consideration of these subtests in the Reading Test.

In terms of developing future versions of the Reading Test part of the UH ELI Placement Test, and other test development projects, recall that the results presented in Tables 5 and 6 were for Random effects models, and they were therefore generalizable to other test projects with the same universe of admissible observations. This means a test should have two facets, items and subtests, and test takers should be expected to answer all items in all subtests (follow Brennan 1983). However, the subtests and items tested could be different. The predicted dependability for the Reading Test reveals that subtests may affect dependability in important ways. Similar findings were reported in Brown's (1996, p. 263) study on the total TOEFL test and its subtests. Therefore, dividing the Reading Test into smaller and smaller subtests is appropriate. However, as also noted by Brown (1996, p. 262), in some cases subtests may have a negligible impact on dependability, which is the case for the RC Test. Another impact on the dependability is from the changing numbers of items, and just as with the subtest facet, its impact varies from case to case. Adding more items to each passage on the reading comprehension subtest (based on the current configuration) appears to be more effective in enhancing the dependability than increasing the number of reading passages with fewer items. On the other hand, increasing numbers of items on each subtest doesn't have much influence on the dependability of the overall Reading Test.

Future Research

During the process of conducting this study, a number of questions emerged. They are presented here as possible topics of future investigations.

1. Would similar results be obtained if the study were replicated with other UH ELI Placement Test data sets? With data from placement tests at other institutions?
2. Since items were not randomly selected for G-sampling, would the result be similar if different items were selected in G-studies?
3. Would the findings be similar for other parts of the test in UH ELI Placement Test?
4. How would the dependability differ at different cut-points?
5. What other methods can be employed to investigate the validity of this test?

REFERENCES

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 239-257.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*(2), 245-258.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brown, J. D. (1984). A norm-referenced engineering reading test. In Pugh, A.K & Ulijin, J.M. (Eds.), *Reading for professional purposes: studies and practices in native and foreign languages.* London: Heinemann.

Brown, J. D. (1990). Short-cut estimators of criterion-referenced test consistency. *Language Testing 7*, 77-97

Brown, J. D. (1993). A comprehensive criterion-referenced testing project. In Douglas, D. & Chapelle, C. (Eds.), *A new decade of language testing research.* Washington DC: TESOL, 163-184.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21-42.

Brown, J. D., & Ross, J. A. (1996). Decision dependability of item types, sections, tests, and the overall TOEFL test battery. In Milanovic, M. & Sville, N. (Eds), *Performance testing, cognition and assessment* (pp. 231-265). Cambridge: Cambridge University Press.

Cronbach, L. J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137-163.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing, 9*(1), 30-49.

Stansfield, C. W., & Kenyon, D. M. (1992). Research of the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20,* 347-364.

Yao Zhang
Department of Second Language Studies
1890 East-West Road
Honolulu, HI 96822

yaozhang@hawaii.edu