Psychological Test and Assessment Modeling, Volume 54, 2012 (3), 211-226

The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test

Hossein Karami[1]

Abstract

This study exploited generalizability theory to explore the impact of persons, items, subtests, and academic background on the dependability of the scores from a high-stakes language proficiency test, the University of Tehran English Proficiency Test (UTEPT). To this end and following Brown (1999), three questions were posed: 1. What are the distributional characteristics and CTT reliability of UTEPT test scores? 2. What are the relative contributions of persons, items, and subtests to the dependability of scores for each group and for all the groups combined? 3. What are the relative contributions of persons, items, subtests, academic background as well as their various interactions to the dependability of the scores when all groups are combined? To investigate the issues, 5795 examinees from four different academic backgrounds were selected from among all the participants who had taken the test in 2004. The results of the study indicated that the relative contributions of the facets were not stable across all groups, though highly similar. In addition, with academic background added as a facet, there was no significant interaction between items and fields, and the dependability of the scores did not decrease either. This result shows that background knowledge does not lead to bias in the UTEPT. This use of G-theory could be extended profitably to other measuring situations.

Key words: Generalizability theory, academic background, UTEPT, dependability, test bias

[1] *Correspondence concerning this article should be addressed to:* Hossein Karami, PhD, Department of English Language and Literature, Faculty of Foreign Language and Literatures, University of Tehran, Kargar Shomali Avenue, Tehran, Postcode/ZIP code: 14155-6553, Iran; email: hkarami@ut.ac.ir

Introduction

It has been long recognized that validity is the single most important consideration in test development and use (Bachman, 1990). Therefore, it is incumbent on test developers and users alike to ensure that the uses to which the tests are put are justified and that test score interpretation is not unduly affected by construct-irrelevant factors (Messick, 1989). Specifically, test scores should be free of bias.

The present paper has exploited Generalizability Theory to investigate the impact of background knowledge on a high stakes language proficiency test, University of Tehran English Proficiency Test (UTEPT). As the test is claimed to be a test of general language proficiency, any impact of background knowledge on test scores would be a case of bias. Therefore, it is of utmost importance to make sure that there is no such impact.

The effect of background knowledge

Content or background knowledge has been shown to impact the performance of test takers on language proficiency tests. In fact, a large body of research studies have been devoted to investigating the issue (e.g. Clapham, 1998; Salmani-Nodoushan, 2002; Krekeler, 2006; Kunnan, 1994; Chihara, *et al*., 1989).

Clapham (1998), for example, conducted a study examining the impact of language proficiency and background knowledge on performance on the Reading Comprehension section of the IELTS (International English Language Testing System) test. The results indicated that background knowledge had an increasing effect on performance as the proficiency levels of the examinees increased. However, she also pointed out that a threshold level of proficiency is needed before test takers are able to take advantage of content familiarity. Furthermore, she hypothesized a threshold for advanced learners above which background knowledge may not be much of an advantage. As she explained, "Above this threshold, readers are so proficient linguistically that they can compensate for a certain lack of background knowledge by making full use of their language resources," (p. 163).

The results of a more recent study by Krekeler (2006) also indicated that background knowledge had significant impact on English for Specific Purposes (ESP) reading tests. Furthermore, he concluded that the majority of examinees "were able to make use of their background knowledge regardless of the level of L2 proficiency" (p. 122).

Other studies have focused on the interaction of items and students with certain academic backgrounds (e.g. Alderson & Urquhart 1985; Hale 1988; Karami 2010; Pae 2004). Pae (2004), for example, undertook a Differential Item Functioning (DIF) study of examinees with different academic backgrounds using Item Response Theory. The examinees belonged to either Humanities or Science group. Pae (2004) reported that seven items were easier for the Humanities, whereas nine items were in favor of the Sciences group. In the Listening part, items which favored the Sciences group dealt with number counting and a

job interview while items favoring Humanities concerned human relationships. For the Reading section, items favoring Sciences pertained to topics such as the story of underwater explorers, data analysis, the story of a fishing village, science and technology, the effect of snow on animals, and sports. On the other hand, items relating to friendship, the life-story of a scientist, and the importance of competition favored the Humanities.

On the other hand, Hale (1988) focused on students from humanities/social sciences vs. students from the biological/physical sciences. The results indicated that there was an interaction effect between the text content and major field on reading test scores. Hale (1988) reported that, "Students in the humanities/social sciences outperformed students in the biological/physical sciences on text related to the former major-field group," (p, 59).

Though revealing in many respects, none of the above studies have focused on the overall impact of background knowledge as a source of variance contributing to error variance. Put another way, to the best of the researcher's knowledge, no study has specifically investigated the effect of background knowledge on the dependability of scores of language proficiency tests. Generalizability theory offers an elegant way of examining such issues and it was exploited in this study.

It should be noted here that in the present study, academic background has been selected to represent the differing content knowledge of different groups of test takers. In the Iranian educational system, high-school students select their majors at the beginning of the second year. There are three majors to be selected: Humanities, Mathematics, and Empirical Sciences. Though the students may change their majors after the high school and select a different major at the university level, this is not frequent. Even if they change their majors after the school, when they sit for the PhD exams, they have gone through different content areas for at least six years. As stated earlier, as the UTEPT is claimed to be a test of general language proficiency, any differential performance of the examinees arising from their different content knowledge will be a source of bias.

Generalizability theory

The major problem with the Classical Test Theory (CTT) based approaches to reliability estimation is that each technique can focus on one and only one source of variance at any one time. Different reliability estimation procedures usually focus on variance caused by such factors as time intervals (test-retest), different formats (equivalent forms), item heterogeneity (internal consistency), and rater inconsistency (mark-remark).

The major contribution of the generalizability theory is that the relative impact of multiple sources of variation on test reliability can be assessed simultaneously (Cardinet, Johnson, & Pini, 2010). Viewed in this way, it is a logical extension of CTT. Apart from the usual sources of variation addressed in CTT, generalizability theory can focus on a variety of test taker characteristics and assess their impact, if any, on test performance.

In contrast to CTT which is based on correlation, generalizability theory rests on ANOVA. As Shavelson, Webb, and Rowley (1989) put it, generalizability theory plays a

similar role in measurement that ANOVA does in research. "Just as the researcher attempts to identify and estimate the effects of potentially important independent variables, G theory attempts to identify and estimate the magnitude of the potentially important sources of error in a measurement" (p. 923).

Under generalizability theory, the performance of any test taker is viewed as a sample from among an infinitely large number of observations that could have been made under similar conditions (Bolus, Hinofotis, & Bailey, 1982). That is, the test users are never interested in the performance of a test taker on this particular set of items, at this time and at this location. Rather, their performance is seen as an exemplar from all possible observations that could be made under similar conditions that the test users would regard as acceptable measures. The most ideal situation would be to observe the performance of test takers under all conditions, with all possible tasks and items, and at all times. However, this is clearly impractical. The best we can do is to identify as many sources of variance as possible and to quantify their impact on test reliability. Generalizability theory provides us with such a mechanism.

The first applications of generalizability theory in language testing began during the 1980s. Bolus *et al*. (1982) provided the first non-technical introduction to the merits of the theory and urged its wider application in language testing. A large number of studies have appeared since then. Lynch and McNamara (1998) applied generalizability theory along with the Many-facet Rasch measurement to examine a test of communicative skills in English as a Second Language for intending immigrants to Australia. Xi (2007) explored the utility of analytic scoring for TOEFL (Test of English as a Foreign Language) Academic Speaking Test (TAST) in providing useful and reliable diagnostic information for operational use in three aspects of candidates' performance: delivery, language use, and topic development.

Kunnan (1992) utilized generalizability theory along with factor and cluster analyses to examine a criterion-referenced test. Bachman, Lynch, and Mason (1995) applied generalizability theory to examine the relative impact of tasks and raters on the dependability of grammar ratings from a Spanish speaking test. Brown (1999) also investigated the relative effect of persons, items, sections, and language background on the paper-based TOEFL using generalizability theory. Brown (1999) also called for further application of the theory to different versions of the TOEFL and also different tests to see whether the results are replicable. Zhang (2006) responded to the call and exploited generalizability theory to investigate the contributions of persons, items, sections, and language backgrounds to the score dependability of the Test of English for International Communication (TOEIC).

The present study is another attempt to estimate the impact of test taker characteristics on the dependability of scores from a language proficiency test. This time, however, instead of focusing on language background, this study focused on background knowledge, here operationalized as academic background.

In Generalizability Theory, bias is defined as an interaction term. That is, if the variance component for the interaction between, say, items and persons is high, it denotes that the difficulty of each item varies across individuals. In other words, some items are easier for

some people and more difficult for others. Hence, it is a case of bias because the items are not performing similarly for all persons. Such an interaction term jeopardizes the generalizability of scores and should be taken into account when interpreting the test results.

The purpose of the study is to examine the following:

1. What are the distributional characteristics and the reliability of the scores obtained from the total test and from each subtest?
2. For each of the four groups with different academic backgrounds, what are the relative contributions of persons, items, and subtests and their interactions? What are their impacts on the dependability of scores?
3. Across all groups, what are the relative contributions of persons, items, subtests, academic background as well as their various interactions? What are their impacts on the dependability of test scores? In particular, is academic background biasing the results of the subtests or of the total test?

Method

Participants

The participants of the present study were a total of 5795 examinees who had taken a version of the University of Tehran English Proficiency Test (UTEPT) in 2004. They were from four different academic backgrounds: Humanities, Science, Technology, and Agriculture. The number of participants in each group is displayed in Table 1. Unfortunately, we did not have access to either the gender or the age of the participants.

Table 1:
Number of participants

	Frequency	**Percent**
Agriculture	776	13.4
Humanities	3368	58.1
Science	767	13.2
Technology	884	15.3
Total	5795	100.0

Materials

The UTEPT test is a language proficiency test which is administered to a large number of PhD applicants each year. The aim of the UTEPT is to identify those individuals who have the right level of English proficiency. The test is composed of three subtests including Grammar and Written Expression, Vocabulary, and Reading Comprehension. The number of questions for each subtest was as follows:

1. Structure and Written Expression (35 items)
2. Vocabulary (25 items)
3. Reading Comprehension (40 items)

All questions were in the multiple choice format. The Reading subtest comprises passages immediately followed by a number of comprehension questions. The number of comprehension questions is different for each passage. Usually, a total raw score is reported to the candidates which is simply the sum of scores they get on the three subtests. All items are dichotomously scored, with correct responses as 1 and incorrect responses as 0. No separate scores are reported for the subscales. All missing responses are also counted as wrong in the scoring of the test.

The test is a high-stakes test because no one would be allowed to sit their respective PhD exams unless they first get the required scores on the UTEPT. As the consequences for the examinees are serious, every attempt should be made to make sure that it is not biased against any group of test takers. This study is one such attempt.

Two-level sampling

The majority of current generalizability analysis software are unable to process incomplete or unbalanced data (Cardinet et al., 2010). Therefore, all missing responses were coded as 0 assuming that the examinees did not know the correct answer. In addition, for the purposes of the study, 3068 participants were randomly selected (767 participants for each group) in order to have equal sized or balanced data for each group. Similarly, as each subtest in the UTEPT does not include the same number of items, we had to randomly select 25 items from the Structure and Written Expression and Reading Comprehension subtests. Thus, each subtest had 25 items. It is clear that after reducing the number of items in these two subtests, the means for each subtest were also reduced. On the other hand, it was made sure that the means and standard deviations were maximally comparable in the original four groups and in the selected four groups of 767 participants.

Administration

Strict conditions have been specified for the administration of the UTEPT. All administration centers begin at the same time and the same amount of time is given to all ex-

aminees. Any mark on the answer key after the standard time is over will count as cheating. No examinee is allowed to bring notes, booklets, pamphlet, or cell phones to the test center. They are not allowed to take notes either.

Analysis

In order to answer the research questions, three steps were followed. First, the distributional characteristics of both the total test and each subtest were obtained for the sample and the original data. In addition, the reliabilities of the total test and of the subtests were checked using Cronbach α, on the basis of the original data.

Second, as per suggestions from Brennan (1983) and following Brown (1999), five generalizability studies were conducted on the data to estimate the relative contributions of persons, items and subtests to the test scores: one analysis consisting of the total data (n=3068) and four separate analyses, one for each group (n=767). The justification for such an analysis is summed up by Brennan (1983): "when a population of objects of measurement is stratified with respect to several clearly defined subpopulations, it is almost always advisable to conduct separate analyses for each subpopulation. In addition, an investigator may want to conduct a global analysis over subpopulations.' (p. 93).

The designs of all the analyses in this part were the same: five two-facet mixed design generalizability analyses were conducted. Persons are the object of measurement and items and subtests are the facets of the study, with persons contributing to the universe score variance and items and subtests contributing to error variance. The persons and items were considered to be a random selection from among an infinite number of items and persons. Thus, they were considered as random facets. On the other hand, the subtests facet was considered as fixed because there were only three subtests in the test and we were not interested in generalizing the results beyond these subtests. The difference between fixed and random facets is that the fixed facets and their interactions do not contribute to measurement error (Cardinet *et al*., 2010). Furthermore, the items were nested within subtests and both were crossed with persons. Therefore, the design of these five analyses was the following: $(i:s)\times p$.

As the items are nested within subtests, the variance component for the items is confounded with the variance component for the items-by-subtests interaction. This leaves five variance components to be estimated: persons, subtests, items nested within subtests confounded with their interaction, the persons-by-subtests interaction, and the variance component for the interaction between items nested within subtests and persons confounded with all undetected variance.

Finally, a single generalizability study was conducted this time adding academic background as a facet. This was a random facet because we intended to generalize the results to other fields. Note also that this time the persons are nested within fields because each person belongs to one and only one field. The design of the study was the following: $(i:s)\times(p:f)$. All the generalizability analyses were conducted using the EduG software (Cardinet, Johnson, & Pini, 2010).

Results and discussion

Score distributions and reliability analyses

The distributions of the scores and the classical reliability analyses are given in Table 2. These pertain to the original data (n=5795), for the total test and each subtest of the UTEPT. (Note that the original data include 100 items, while the G-studies to be presented later are based on 75 randomly selected items.) The estimates of the reliabilities for the real unabridged UTEPT, computed for the total sample of participants (n=5795), are .90, .80, .77, and .78 for the Total test, Grammar, Vocabulary, and Reading subtests respectively. A word of caution is in order. Cronbach alpha assumes that the items are drawn at random from a common pool of items. This assumption is violated when a test is comprised of a number of subtests, each measuring a different aspect of the construct, as is the case with the UTEPT. Therefore, the reliability estimate of 0.90 reported here for the UTEPT Total score should be taken as a lower bound approximation of the real reliability index.

Table 2:
Descriptive statistics and reliability analyses

Test Subtest	Mean	SD	Number of items	Number of persons	Cronbach α
UTEPT Total	49.25	14.221	100	5795	.90
Grammar	19.44	5.954	35	5795	.80
Vocabulary	13.24	4.629	25	5795	.77
Reading	16.57	6.127	40	5795	.78
UTEPT Sample	37.82	10.264	75	3068	.85
Grammar	13.92	4.034	25	3068	.70
Vocabulary	13.60	4.354	25	3068	.74
Reading	10.3061	4.00666	25	3068	.67

Five generalizability studies (the $i{:}s \times p$ design)

In order to answer the second research question and examine the relative contributions of persons, items, and subtests, five generalizability studies were conducted: one overall study including all participants (3068) and four separate studies for each academic background each study including 767 participants.

The results of these five generalizability studies are summarized in table 3. The important point to be noted here is that the design of all these studies is the same and that it is a mixed design where one of the facets, the subtests, is fixed. The problem with mixed

designs is that the mean squares for the random facets are calculated by dividing the total sums of squares by the degree of freedom, that is N-1. On the other hand, the mean squares for a fixed facet comes about by dividing the total sums of squares by the sample size, N. This happens due to the fact that in the case of fixed facets, we don't have sampling error because no sampling is done at all. Clearly, this is a hurdle in the way of the comparability of the two estimates. EduG reports a corrected variance component which is based on a correction formula proposed by Whimbey, Vaughan, & Tatsuoka (1967 as cited in Cardinet *et al*., 2010). The correction proceeds by multiplying the result of the ANOVA by (*N* - 1)/*N*. Note how the correction depends on the size of the population. Whenever the size of the population tends towards infinity, the correction has virtually no impact on the results. For small population or universe sizes, however, the correction can make a large difference. The variance components reported here are the corrected estimates.

As is evident from table 3, the most significant variance component in all studies is the interaction between persons and items nested within subtests. This variance component is also compounded with all undetected sources of variation and/or random error variance.

The next highest variance component is the I:S, items nested within subtests. The percentages of this variance range from 7.4 for Humanities to 11.2 for Technology. This shows that items were not of the same difficulty level. This difference in difficulty level was most significant in the Technology group and least important for the Humanities.

The next variance component of interest here is that for persons. This shows how much the persons are spread out by the test. That is, how much the test can differentiate between examinees. Again, not all generalizability studies display the same variance component for this facet. The percentages range from 5.1 for Agriculture to 7.9 for Humanities.

Finally, two sets of coefficients are reported for each G-study: relative G coefficient and absolute G coefficient. These two statistics are suitable for different decisions. If the test is norm-referenced, the relative G coefficient is the important one because it shows how much the relative standing of the individuals are generalizable. Here, the performance of the individual is gauged with respect to his relative standing among a group of test takers. On the other hand, if the test is criterion-referenced, then the absolute G coefficient is of interest. In this case, we are interested in the performance of the examinee alone with no importance attached to his relative standing. The absolute coefficient is usually lower than the relative one because more facets contribute to absolute error variance.

These G coefficients are estimated in a similar way to reliability estimation in CTT. In both cases, the true variance is divided by the total variance. In these five studies, the variance component contributing to universe score variance is the persons facet. The higher the variance component for this facet, the higher the reliability of the test. It was noted earlier that the generalizability study for the Humanities group had the highest variance component for persons. As the error variances are almost the same across the groups, the highest G coefficients for the Humanities is due to the high variance component for persons in this group. By the same token, the Agriculture group had the least

Table 3:
G study results for the four groups and the total

	Total			Humanities			Science			Technology			Agriculture		
Source	SS	VC	%	SS	VC	%	SS	VC	%	SS	VC	%	SS	VC	%
P	4394.18416	0.01638	6.5	1302.35247	0.01995	7.9	905.47442	0.01303	5.2	1072.98260	0.01606	6.5	893.47984	0.01280	5.1
S	986.43335	0.00367	1.5	217.44525	0.00327	1.3	258.19122	0.00379	1.5	298.15472	0.00443	1.8	223.38611	0.00323	1.3
I:S	5079.72313	0.02293	9.1	1041.65977	0.01860	7.4	1445.97246	0.02592	10.3	1554.71979	0.02790	11.2	1331.77992	0.02385	9.5
PS	2008.01998	0.00493	2.0	516.95475	0.00534	2.1	487.70211	0.00455	1.8	446.51195	0.00379	1.5	546.10722	0.00600	2.4
PI:S	45051.95687	0.20402	81.0	11245.54023	0.20390	81.2	11283.46754	0.20459	81.2	10844.56021	0.19663	79.0	11383.98008	0.20641	81.8
Total	57520.31750		100	14323.95247		100	14380.80775		100	14216.92927		100	14378.73318		100
Coef_G relative	0.86			0.88			0.83			0.86			0.82		
Coef_G absolute	0.84			0.87			0.81			0.84			0.80		
Relative ErrV	0.00272			0.00272			0.00273			0.00262			0.00275		
Absolute ErrV	0.00303			0.00297			0.00307			0.00299			0.00307		
Relative SE	0.05216			0.05214			0.05223			0.05120			0.05246		
AbsoluteSE	0.05501			0.05447			0.05544			0.05471			0.05541		

Coef_G relative: G-coefficient for relative decisions
Coef_G absolute: G-coefficient for absolute decisions
Relative ErrV : Relative Error Variance
Absolute ErrV : Absolute Error Variance

variance component for persons. It is evident in the lowest G coefficient the study for this group displayed.

In sum, it is clear that the relative contributions of different variance sources are not the same across all groups. The variance components for I:S, items nested within subtests, range from 7.4 for Humanities to 11.2 for Technology. Furthermore, the variance components for persons range from 5.1 for Agriculture to 7.9 for Humanities. Although the variance components are not completely uniform across the groups, it should be borne in mind that they are similar. Specially, the order of the components is uniform across the groups. This may indicate that the small differences in the distribution of variance components may not be of much significance. Finally, the G coefficients are higher than .8 in all G-studies.

Single combined generalizability study (the $p{:}f \times i{:}s$ design)

The final generalizability study included all the facets in a single design: $p : f \times i : s$. The number of participants for this study was 3060. The design includes persons nested within fields, fields, items nested within subtests and the subtests plus all the interactions between these facets. Note that the subtests facet is again fixed. Therefore, neither the main effect for this facet nor its interactions with other facets contribute to error variance.

The results of the single overall analysis of variance are presented in Table 4. By far the largest variance component is the interaction between persons, items, subtests, fields and other undetected sources of variation or just random variance.

Probably the most significant part of the table is the large variance component for the main effect of items or I:S. Such a large variance component for this main effect shows that the items are not of the same difficulty level. Such a large variance component increases the difficulty of generalizing from a student's observed score on this test to his universe score and adds to absolute measurement error (Shavelson and Webb,1991). This variance component does not affect relative measurement error because it is constant for all examinees. It is also evident from Table 4 that the next largest variance component is due to the persons or P:F. It is clear that the persons were not of the same ability levels. Such a large variance component for persons adds to the reliability or generalizability of the scores and is, therefore, desirable.

Note also that the main effect for subtests is not zero. This indicates that subtests were not of the same difficulty levels. Furthermore, the variance component for fields is close to zero. It is an indication of the fact that the examinees from different backgrounds are, on the whole, of similar ability levels. That is, the overall performances of the groups are similar. It is also evident from this table that the interaction between field and subtests is near 0. It shows that the relative difficulty level of subtests remains stable across all fields.

Table 4:
Analysis of Variance

Source	SS	VC	%
F	219.89483	0.00092	0.4
P:F	4174.28934	0.01546	6.1
S	986.43335	0.00367	1.5
I:S	5079.72313	0.02293	9.1
FS	10.74395	0.00001	0.0
FI:S	294.40881	0.00151	0.6
PS:F	1997.27604	0.00492	2.0
PI:FS	44757.54806	0.20288	80.4
Total	57520.31750		100%

The main objective of the present study was to study the impact of various factors, but especially of the participants' field of study, on the test's fairness. In other words, the study aimed essentially at checking that no bias existed in favor or against students from different academic backgrounds. As stated earlier, in G-theory, bias can be detected when an important interaction is found between the object of the study and another facet (here, academic background).

The relative contributions of the various facets in the $p:f\times i:s$ design and their interactions to the relative and absolute error variance along with relative and absolute G coefficients are reported in Table 5.

The differentiation variance has two sources: variation between fields of study F and variation between participants within each field of study P:F. The effect of the fields is small, since its variance represents less than one tenth of the variance between participants. But added together, these two variances explain approximately 85% of the total observed variance of the examinees' total scores. The reliability of a test formed with these selected items could be estimated as 0.85. But this value is smaller than the real reliability, since the real test UTEPT is longer than the truncated samples of items used in the G-study.

The components of the absolute error variance are the subtests S and the items nested in the subtests (I:S). The facet S being fixed does not contribute to the error variance, but the varying difficulty of the items in each subtest causes 10% of the absolute error variance.

The relative error variance has four components. The first two are directly related to the fields of study: the interaction between Fields and Subtests FS, and the interaction between Fields and Items within Subtests FI:S. As the facet Subtests S is fixed, it does not cause sampling variations. Thus, it is not a source of measurement error.

Table 5:
Contributions to relative and absolute error variance of different facets

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
F	0.00092		.....		.....	
P:F	0.01546		.....		.....	
	.....	S	.....		(0.00000)	0.0
	.....	I:S	.....		0.00031	10.1
	.....	FS	(0.00000)	0.0	(0.00000)	0.0
	.....	FI:S	0.00002	0.7	0.00002	0.7
	.....	PS:F	(0.00000)	0.0	(0.00000)	0.0
	.....	PI:FS	0.00271	99.3	0.00271	89.2
Sum of variances	0.01638		0.00273	100%	0.00303	100%
Standard deviation	0.12799		Relative SE:	0.05220	Absolute SE:	0.05505
Coef_G relative	0.86					
Coef_G absolute	0.84					

The interaction between Fields and Items within Subtests FI:S, on the contrary, could be a source of bias. But its value is quite small, less than 1% of the relative or absolute error variances. Therefore, it may be safely concluded that the UTEPT is not biased against any group.

The last two components of the relative error variance are the interactions of persons in fields with subtests PS:F and the interactions of persons in fields with items nested in subtests PI:FS. As subtests are fixed, their interaction with persons in fields PS:F does not contribute to measurement error. On the contrary, the last source of error, PI:FS, the interactions of persons in fields with items nested in subtests, is the essential cause of measurement error, representing 89% of the absolute error variance. However, as it results from the random encounter of one particular person with one particular item, it cannot easily be reduced, except by increasing the number of items.

Generally, the minimum acceptable value for the G-coefficients is considered to be .80 (Cardinet et al., 2010). The G-Coefficients we have come up with in this study, .86 and .84 are higher than the standard required level. It is clear that the UTEPT enjoys a high level of reliability. Although the coefficients reported here are based on an analysis of a

reduced number of items and may not completely mirror those of the complete test, they are still above the minimum acceptable level.

Conclusion

The results of both CTT reliability analyses and generalizability studies indicated that the UTEPT enjoys an acceptable reliability level. Both relative and absolute G coefficients for all G studies were above 0.8. The global generalizability study with fields included as a facet resulted in G coefficients of .86 and .84 for both relative and absolute decisions. This is clearly a favorable situation bearing in mind the fact that these G coefficients are based on a truncated version of the test and that they are lower-bound estimates of the real G coefficients.

The five G-studies indicated that the variance components for different facets and their interactions varied from group to group. For all groups, however, the largest variance component corresponded to the interaction of highest level, the one which is confounded with the uncontrolled sources of variation. The Items facet (I:S) had the second highest variance component. This indicates that the test items in each subtest were of varying difficulty levels. For all groups also, the Persons facet (P:F) produced a relatively high variance component, but smaller that for the Items. That fact reduces the dependability of the total test score.

The final global generalizability study, with fields added as a facet, did not display a high variance component for the item by field interaction. This indicated that the relative difficulty levels of the items are stable across all fields. It is clearly an indication of lack of bias due to items in the test. The main effect for fields was also close to zero which shows that the mean abilities of examinees from different fields were relatively similar.

Therefore, it may be concluded from the results of the present study that academic background did not exert a large influence on the performance of the examinees on the UTEPT. Therefore, this test may be considered unbiased against any academic groupings. Of course, such interpretations should be made cautiously due to the scope of the present study. Further research will undoubtedly shed more light on this issue.

Finally, further research is clearly needed before firm conclusions can be made about the role of academic background on score dependability. The research may focus on different tests or different versions of the UTEPT test examined in this study.

References

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*(2), 192-204.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-57.

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

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing,16*, 217-38.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982) An introduction to generalizability theory in second language research. *Language Learning*, *32*(2), 245-58.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York: Routledge.

Chihara, T., Sakurai, T., & Oller, J. W. (1989). Background and culture as factors in EFL reading comprehension. *Language Testing, 6*(2), 143-151.

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 141-168). Mahwah, NJ: Lawrence Erlbaum.

Karami, H. (2010). *A differential item functioning analysis of a language proficiency test: an investigation of background knowledge bias.* Unpublished MA Thesis, University of Tehran.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. *Language Testing, 23*(1), 99-130.

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.

Kunnan, A. J. (1994). Modeling relationships among some test-taker characteristics and performance on EFL tests: an approach to construct validation. *Language Testing, 11*(3), 225-52.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-Facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education & Macmillan.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing, 21*(1), 53-73.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and psychological Measurement, 59*(2), 248-70.

Salmani-Nodoushan, M. A. (2002). *Text familiarity, reading tasks, and ESP test performance: a study on Iranian LEP and non-LEP university students.* Unpublished PhD dissertation, University of Tehran.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability Theory. *American Psychologist, 44*(6), 922-32.

Swiss Society for Research in Education Working Group (2010). *EDUG user guide*. Neuchatel, Switzerland: IRDP.

Whimbey, A., Vaughan, G. M., & Tatsuoka, M. M. (1967). Fixed effects vs. random effects: estimating variance components from mean squares. *Perceptual and Motor Skills*, *25*, 668.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2), 251-86.

Zhang, S. (2006) Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing, 23*(3), 351-69.