*International Journal of Assessment and Evaluation in Education*
*Vol 1/Dec 2011 (1-37) ISSN 2232-1926*

# What do the L2 Generalizability Studies Tell Us?[1]

**James Dean Brown**
*University of Hawai'i at Manoa*
*2500 Campus Rd, Honolulu, HI 96822,*
*United State of America*

## *Abstract*

*This research synthesis examines the relative magnitudes of the variance components found in 44 generalizability (G) theory studies in L2 testing. I begin by explaining what G theory is and how it works. In the process, I explain the diffrences between relative and absolute decisions, between crossed and nested facets, and between random and fixed facets, as well as what variance components (VCs) are and how VCs are calculated. Next, I provide an overview of G-theory studies in L2 testing and discuss the purposes of this research synthesis. In the methods section, I describe the materials used in this research synthesis in terms of the samples of students, the tests, and the G-study designs used. I also present the analyses in terms of how the data were compiled and analyzed. The results are sorted and displayed to reveal patterns in the relative contributions to test variance of various individual facets as well as interactions between and among facets for different types of tests. I next discuss these patterns and put them into perspective. I conclude by exploring what I think the results mean for L2 testing in general.*

***Keywords*** *generalizability theory, norm-referenced relative decisions, measurement facets, variance components*

## INTRODUCTION

### Generalizability Theory

Cronbach, Rajaratnam, and Gleser (1963) first proposed *generalizability theory* as a useful extension of classical theory reliability (to review classical theory reliability, see Bachman, 2004, pp. 153-191; Brown, 2005a, pp. 169-198). Generalizability (G) theory takes reliability to be a question of the degree to which one can generalize from one observation to a universe of observations (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 15). G theory then allows generalizing from a specific sample to the universe of interest by means of a set of clearly defined estimation procedures (Shavelson & Webb, 1981, pp. 133 – 137). Using analysis of variance (ANOVA) procedures, testers can segregate and estimate the relative magnitude of variance components (VCs) associated with various *measurement facets* in a G study (Suen,

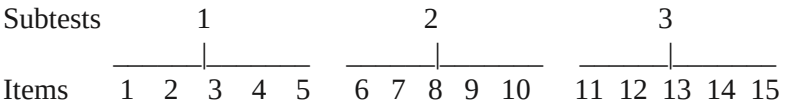
1990, pp. 41 – 42). Based on the estimated VCs, the researcher can further study how various potential modifications of measurement facets will likely affect the generalizability coefficient (analogous to a reliability coefficient) of the test. Testers can then make test design decisions that are based on more accurate estimates of the effects of error than were previously available in classical theory. Shavelson and Webb (1991), Brennan (1983, 2001), and Chiu (2001) explain these G-study estimation procedures in more detail.

## Relative decisions versus absolute decisions

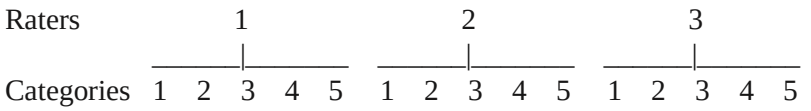
One important feature of G theory is that it can account for differences in the dependability of *norm-referenced* tests, which are used to make what are called *relative decisions* in G theory, and *criterion-referenced* tests, which are used to make *absolute decisions*. This is accomplished by defining which sources of error are included in the calculations. Shavelson and Webb (1991) and Brennan (1983, 2001) provide particularly useful explanations for these estimation procedures for both norm-referenced relative decisions and criterion-referenced absolute decisions. In this research synthesis, I will examine only those tests designed for norm-referenced relative decisions.

### Crossed facets versus nested facets

In looking at measurement facets, one source of confusion that arises is the difference between crossed facets and nested facets. The first source of confusion arises because these are not labels for the facets themselves, but rather for the relationships between facets. For example, items are said to be *nested* within subtests if there are different items in each subtest. This is typically the case for the subtests on multiple-choice standardized tests. For example, a reading comprehension test with three passages (subtests) might have five items on each passage, but they would naturally be *different* items for each passage. Thus items would be *nested* within subtests (symbolized as i:s). The fact that items are *nested* within subtests describes the relationship between items and subtests, which is *nested* because the members of the subordinate category (items) are *different* for each of the subtests as follows:




In contrast, if three raters were using the *same* five categories (say content, organization, mechanics, language use, and vocabulary) to rate a series writing samples, the categories would be *crossed* with raters (symbolized as cr). The fact that categories are crossed with subtests describes the relationship between categories and subtests, which is *crossed* because the members of the subordinate categories are the *same* for each rater as follows:




Examinees are most often called *persons* in G studies. If persons were crossed with the other facets because all the persons were tested under the same conditions (i.e., all persons took the same fifteen items in the same three subtests, or all persons were rated in the same five categories by the same three raters), the first G-study design above would be referred to as a persons crossed with items nested within subtests (pi:s) design, while the second G-study design above would be called a persons crossed with categories crossed with raters (pcr) design.

**Random versus fixed facets**

In designing the analyses for a G study, researchers must consider whether each of their facets is random or fixed. The choice revolves around whether the sample for each facet was a *random* sample from the universe of all possibilities (or can at least be considered *exchangeable* with any other similarly sized sample of all possibilities) or was fixed in the sense that the sample "exhausts the conditions in the universe to which the researchers want to generalize" (Shaveleson & Webb, 1991, pp. 11–12). The choice is sometimes determined by the nature of the facet, but can also be decided rather arbitrarily by the researchers based on whether or not they buy into the notion of exchangeability mentioned in the previous sentence.

**What are Variance Components?**

The first stage of a G-theory project is a G study. Based on the mean squares obtained in an ANOVA procedure, *variance components* (VCs) are estimated for different sources of variation, that is, for main effects and interactions across the object of measurement and the facets of measurement the researcher chooses to model. Consider for example a G study designed to investigate the relative effects of the object of measurement (persons) crossed with the main effects (for items nested within subtests, or pi:s, which are the facets of measurement modeled by the researchers) and their interactions. The first step is to conduct an ANOVA as shown in Table 1 (based on the data used in Brown & Ross, 1996).

**Table 1** ANOVA for a pi:s Design (adapted from Brown & Ross, 1999)

| **SOURCE** | **SS** | **df** | **MS** |
|---|---|---|---|
| persons (p) | 80306.73 | 19999 | 4.0155373 |
| subtests (s) | 4200.90 | 2 | 2100.4500000 |
| items nested in subtests (i:s) | 24395.20 | 111 | 219.7765766 |
| persons by subtests (ps) | 17417.30 | 39998 | .4354543 |
| persons by items nested within subtests (pi:s) | 359188.37 | 2219889 | .1618047 |

Using what is known about how VCs make up the estimated mean squares (*EMS*) (as shown in Kirk, 1968; Brennan, 1983; Brennan, 2001)[2], the *EMS* formulas given in

the third column of Table 2 are used as shown in the fourth column to calculate the VCs from the observed mean squares (*MS*) for the p, s, and i:s, as well as for the ps and pi:s interactions.

**Table 2** EMS Used to Derive VC in a pi:s Design (adapted from Brown & Ross, 1999)

| SOURCE | MS | EMS | Calculating VC | VC |
|---|---|---|---|---|
| p | 4.01553728 | $\sigma^2(pi:s) + n_i\sigma^2(ps) + n_in_s\sigma^2(p)$ | $(MS_p\text{-}MS_{ps})/n_in_s$ | .03140424 |
| s | 2100.45000000 | $\sigma^2(pi:s) + n_i\sigma^2(ps) + n_p\sigma^2(i:s) + n_pn_i\sigma^2(s)$ | $(MS_s\text{-}MS_{i:s}\text{-}MS_{ps}+MS_{pi:s})/n_pn_i$ | .00247421 |
| i:s | 219.77657658 | $\sigma^2(pi:s) + n_p\sigma^2(i:s)$ | $(MS_{i:s}\text{-}MS_{pi:s})/n_p$ | .01098074 |
| ps | .43545427 | $\sigma^2(pi:s) + n_i\sigma^2(ps)$ | $(MS_{ps}\text{-}MS_{pi:s})/n_i$ | .00720131 |
| pi:s | .16180465 | $\sigma^2(pi:s)$ | $MS_{pi:s}$ | .16180465 |

Given the *EMS* shown in Table 2, the VCs can be systematically derived from the *MS*. Starting at the bottom of Table 2 and working up through the formulas in Table 2, the five VCs in this design can be calculated[3] as follows:

$\sigma^2(pi:s) = MS_{pi:s} = .16180465$

$\sigma^2(ps) = (MS_{ps}\text{-}MS_{pi:s})/n_i = (.43545427 - .16180465)/38 = .00720131$ $\sigma^2(i:s) = (MS_{i:s}\text{-}MS_{pi:s})/n_p = (219.77657658 - .16180465) / 20000 = .01098074$ $\sigma^2(s) = (MS_s\text{-}MS_{i:s}\text{-}MS_{ps}+MS_{pi:s})/n_pn_i = (2100.45000000 - 219.77657658 - .43545427 + .16180465) / (20000 \times 38) = 1880.3997738 / 760000 = .00247421$

$\sigma^2(p) = (MS_p\text{-}MS_{ps})/n_in_s = (4.01553728 - 0.43545427)/(38 \times 3)$
$= 3.58008301 / 114 = 0.03140424$

At the end of this G-study stage, it only remains for the VCs for the object of measurement, each facet, and the appropriate interactions to be interpreted in terms of their relative magnitude. It is important to recognize that *each of these VCs represents the results for a single observation.* Thus the VC for persons of .03140424 represents the variance for a *single* person, the VC for subtests of .00247421 is for a *single* subtest, and so forth.

Once the G study is finished, the researcher then shifts to a second stage, called a decision study (D study), in which the G-study VCs are used to further calculate generalizability indices, signal to noise ratios, and/or phi coefficients, for different testing purpose and various combinations of numbers of facets. For instance, in the running example here, with its pi:s design, a researcher might want to calculate generalizability coefficients (analogous to classical test theory reliability estimates) for relative decisions or absolute decisions (depending on whether the purpose of the test is norm-referenced or criterion-referenced, respectively), and for various numbers of items and subtests. The goal of the D study is to explore different possible test designs to see which might be best for a revised version of the test in terms of overall generalizability given the testing purposes and practical constraints (e.g., on item and

subtest writing), but also given things like time allowed for the test administration, student fatigue, etc.

## Generalizability Theory in Language Testing[4]

In the process of gathering studies for this research synthesis, I searched for G studies in book chapters, dissertations, conference presentations, and journal articles. Naturally, I used my experience in doing the literature reviews for my own G studies as well as Google Scholar to identify many G studies in books and conference presentations. I also systematically searched UMI Dissertation Abstracts online for related dissertations as well as all the research reports available on the Educational Testing Service website. The journals I examined were all issues of *Language Testing*, *Language Assessment Quarterly*, and *International Journal of Testing*. I also searched those issues electronically available to me (which were therefore published in the last decade or so) of *Applied Measurement in Education*, *Applied Psychological Measurement*, and *Journal of Educational Measurement*. I found no second/foreign language testing G studies in any of these mainstream measurement journals. If I missed appropriate G studies in earlier issues of those mainstream journals or in other mainstream journals that I did not survey, that contributes bias into my selection process.

For a G study to be included in this research synthesis it had to: (a) be based on second or foreign language testing, (b) have *persons*, pure and simple (i.e., not in a nested relationship with any other facet), as the primary object of measurement, (c) be norm-referenced in focus, and (d) present a complete set of raw VCs for the G studies in question.

## G-theory papers in the language testing not directly applicable to the present paper

The idea of applying G theory to language testing first surfaced in Bolus, Hinofotis, and
Bailey (1982), but it was not actually applied in that paper. Others have also described G theory without applying it. For instance, Bachman (1997) briefly described G theory in terms of the concepts, procedures, problems and solutions, and suggestions for future research. Brown and Hudson (2002, pp. 184–197) discussed applications of G theory to criterion-referenced language testing, and Bachman (2004, pp. 176–188) briefly introduced some of the key concepts in G theory.

Other authors have applied G theory but for purposes not directly related to the present research synthesis. For instance, Bachman, Lynch, and Mason (1995), Brown (1990b, 1993), Kunnan (1992), and Sawaki (2003, 2007) used G theory to analyze criterion-referenced tests. Five studies included facets not found in any of the other studies: Stansfield and Kenyon (1992) used G theory to compare *testing formats* (OPI vs. SOPI); Van Moere (2006) included testing *occasions* as a facet; Abeywickrama (2007) used G theory to compare *tests types* (cohesion and coherence); Kim 2009 focused on *rater types* (native-speaker vs. non-native speaker); and Gerbil (2009, 2010) focused on *tasks types* (independent vs. integrated). These studies were eliminated because no patterns for *testing formats*, *occasions*, *test types*, *rater types*,

or *task types* could emerge from including them. Yamamori (2003) used G theory to evaluate the generalizability of students' interest, willingness, and attitude toward English lessons. In two other papers (Molloy & Shimura, 2005; Gao & Rodgers, 2007), persons were not the object of measurement. Brown (2005b) discussed G theory and decision studies with specific reference to the Malloy and Shimura (2005) paper. For their own reasons, five studies (Lee, Gentile, & Kantor, 2008; Park, 2007; Schoonen, 2005; Xi, 2007; and Xi & Mollaun, 2006) provided VCs for the individual rating categories being measured in their studies, but not for the overall tests (i.e., for persons, raters, and categories combined); hence, their results, while interesting, were not directly comparable to those of the other studies included here. And finally, Sawaki (2003, 2007) reported her D-study results, but not the basic G-study VCs. Indeed, the results in all the studies discussed in this paragraph were interesting, but not directly applicable or relevant to the present paper.

### *Directly applicable G-theory studies in language testing*

Table 3 summarizes the facets included in the 44 G studies relevant to the present paper. Note that, in a number of cases, the different G studies in this research synthesis are based on the same data. For example, there are five rows in Table 3 for Brown and Ross (1993). While these five rows represent different G-study designs, they were applied to different combinations of the same data. Thus there are dependencies among the studies that the reader must keep in mind. This is also true for Zhang (2003), Yoshida (2004, 2006), Zhang (2004, 2006), Yamanaka (2005), Alharby (2006), Tang (2006), and Brown (2008). The different G study designs are nonetheless interesting because of what they reveal differentially in the patterns of VC percentages discussed below in the results.

Notice that Table 3 describes each G study in terms of the author(s) and date(s) of publication, the original facet labels used in those studies, as well as the numbers of persons, items, subtests, categories, (item) types, and raters. Notice also that Table 3 indicates that all of these studies were balanced designs in the sense that all levels of all facets had the same number of observations. For the purposes of the present research synthesis, whether they were labeled *students*, *examinees*, *testees*, etc. in the original papers, such people are referred to by the more traditional G-theory label of *persons*. In addition, *tasks* and *items* are both referred to as *items*, *raters* and *ratings* are referred to as *raters*,[5] and *subsections*, *passages, text types*, *functions*, *subskills*, and *subtests* are all labeled as *subtests* (as will become clear in the Results section, the distinctions among these different types of labels for what I am collapsing into *persons*, *items, raters,* and *subtests* made little difference to the patterns found in this research synthesis).

**Table 3** Summary of Facets and Designs in the G Studies Examined in the Present Research Synthesis

| Author(s) & Date(s) | Original Facet Labels (below) Labels in this study à | Persons | Languages | Items | Subtests | Categories | Item) Types | Raters | Design* |
|---|---|---|---|---|---|---|---|---|---|
| Brown, 1982, 1984 | Persons x Items:Passages | 78 | | 60 | 3 | | | | pi:s |
| Brown & Bailey, 1984 | Persons x Categories x Raters | 50 | | | | 5 | | 10 | pcr |
| van Weeren & Theunissen, 1987 | Testees x Items x Raters | 26 | | 24 | | | | 14 | pir |
| van Weeren & Theunissen, 1987 | Testees x Items x Raters | 29 | | 40 | | | | 15 | pir |
| Brown & Ross, 1993 | Persons x Items:Subsections | 20,000 | | 114 | 3 | | | | pi:s |
| Brown & Ross, 1993 | Persons x Items:Subsections | 20,000 | | 45 | 3 | | | | pi:s |
| Brown & Ross, 1993 | Persons x Items:Subsections | 20,000 | | 28 | 2 | | | | pi:s |
| Brown & Ross, 1993 | Persons x Items:Subsections | 20,000 | | 58 | 2 | | | | pi:s |
| Brown & Ross, 1993 | Persons x Items:Subsections | 20,000 | | 20 | 5 | | | | pi:s |
| Lynch & McNamara, 1998 | Persons x Items x Raters | 83 | | 23 | | | | 4 | pir |
| Brown, 1999 | Persons x Items:Subtests | 15,000 | | 114 | 3 | | | | pi:s |
| Shin, 2002 | Persons x Items:Types:Subtests | 157 | | 48 | 3 | | 4 | | pi:t:s |
| Xi, 2003 | Persons x Tasks x Raters | 20 | | 12 | | | | 4 | pir |
| Zhang, 2003 | Persons x Items:Subtests | 94 | | 75 | 3 | | | | pi:s |
| Zhang, 2003 | Persons x Items:Passages | 94 | | 24 | 6 | | | | pi:s |
| Zhang, 2003 | Persons x Items | 94 | | 50 | | | | | pi |
| Zhang, 2003 | Persons x Items | 94 | | 25 | | | | | pi |
| Zhang, 2003 | Persons x Items | 94 | | 25 | | | | | pi |
| Kozaki, 2004 | Persons x Tasks x Items x Judges | 20 | | 4 | | 7 | | 4 | picr |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Yoshida, 2004, 2006 | Persons x Items:Categories x Raters | 60 | | | 15 | 6 | pi:cr |
| Yoshida, 2004, 2006 | Persons x Items:Categories x Raters | 60 | | | 15 | 6 | pi:cr |
| Zhang, 2004, 2006 | Persons x Items:Sections | 90,312 | 200 | 2 | | | pi:s |
| Zhang, 2004, 2006 | Persons x Items:Sections | 45,156 | 200 | 2 | | | pi:s |
| Zhang, 2004, 2006 | Persons x Items:Sections | 45,156 | 200 | 2 | | | pi:s |
| Yamanaka, 2005 | Persons x Rating Scale Items x Raters | 20 | | | 5 | 10 | pcr |
| Yamanaka, 2005 | Persons x Rating Scale Items x Raters | 20 | | | 5 | 6 | pcr |
| Yamanaka, 2005 | Persons x Rating Scale Items x Raters | 20 | | | 6 | 10 | pcr |
| Yamanaka, 2005 | Persons x Rating Scale Items x Raters | 20 | | | 6 | 6 | pcr |

**Table 3** *(cont.)*

| **Author(s) & Date(s)** | **Original Facet Labels (below) Labels in this study à** | **Persons** | **Languages** | **Items** | **Subtests** | **Categories** | **Item) Types** | **Raters** | **Design*** |
|---|---|---|---|---|---|---|---|---|---|
| Lee, 2005, 2006 | Persons x Tasks x Raters | 261 | | 11 | | | | 2 | pir |
| Lee & Kantor, 2005, 2007 | Persons x Tasks x Raters | 488 | | 6 | | | | 2 | pir |
| Alharby, 2006 | Persons x Raters | 233 | | | | | | 4 | pr |
| Alharby, 2006 | Persons x Raters | 233 | | | | | | 4 | pr |
| Alharby, 2006 | Persons x Domains x Raters | 233 | | | | 4 | | 4 | pcr |
| Alharby, 2006 | Persons x Domains x Raters | 233 | | | | 4 | | 4 | pcr |
| Tang, 2006 | Students x Tasks x Ratings | 9351 | | 2 | | | | 3 | pir |
| Tang, 2006 | Students x Tasks x Ratings | 6818 | | 2 | | | | 3 | pir |
| Tang, 2006 | Students x Tasks x Ratings | 3243 | | 2 | | | | 3 | pir |

| | | | | | | |
|---|---|---|---|---|---|---|
| Tang, 2006 | Students x Tasks x Ratings | 1099 | 2 | | 3 | pir |
| Banno, 2008 | Candidates x Tasks x Raters | 6 | 3 | | 61 | pir |
| Banno, 2008 | Candidates x Tasks x Raters | 6 | 3 | | 61 | pir |
| Brown, 2008 | Persons x Items:Functions x Raters | 53 | 24 | 3 | 4 | pi:sr |
| Brown, 2008 | Persons x Items:Functions x Raters | 53 | 24 | 3 | 4 | pi:sr |
| Brown, 2008 | Persons x Items:Functions x Raters | 53 | 8 | 3 | 4 | pi:sr |
| Brown, 2008 | Persons x Items:Functions | 53 | 8 | 3 | | pi:s |

* p = persons; I = items or tasks; s = subtests; r = raters; c = categories; t = (item) type

## PURPOSE

Glass defined *meta-analysis* as early as (1976, p. 3) as "…the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings." Meta-analyses in second language studies have been conducted on a number of topics (e.g., Sahari, 1997; Ross, 1998; Blok, 1999; Norris & Ortega, 2000; Godschneider & deKeyser, 2001; Masgoret & Gardner, 2003; Rolstad, Mahoney, & Glass, 2005; Russell & Spada, 2006; Jeon & Kaya, 2006; Taylor, Stevens, & Asher, 2006; and Mackey & Goo, 2007) (for an introduction/overview of meta-analysis in second language research, see Oswald & Plonsky, 2010).

This paper is not a meta-analysis, but is rather a *research synthesis* in that the analysis: (a) is based on carefully rationalized selection of G studies; (b) examines the actual data reported in each of the G studies (the VCs in this case), not what the researchers say or claim they found; and (c) uses a clear protocol that delineates what should be consistently and thoroughly considered in each study and across studies (Norris & Ortega, 2006, pp. 807-808). As will be explained below, the present study has all three characteristics. Thus it is a research synthesis, and makes no claims to being a meta-analysis because it does not include statistical analysis of the aggregated results (for more on research synthesis in second language studies, see Norris & Ortega, 2006, 2007).

In this research synthesis, I report the VCs, but I focus on the percentages of variance accounted for by each of the VCs within each study because the relative importance of the VCs for each facet is then comparable across studies as well. To do this properly, I have gone back over the VCs, converted them into percentages, and attempted to understand them in terms of each study, as well as in terms of what the patterns of VC percentages across the many G studies can tell us.

More formally, G theory has been used widely in language testing for a variety of purposes. The purpose of this research synthesis is to compare and contrast the G-

study VCs obtained in a wide variety of language testing studies in order to address the following five research questions:

1. What are the relative magnitudes of the main-effects VCs reported in G studies of the multiple-choice tests for with persons, items, and subtests?
2. What are the relative magnitudes of the VCs for interactions reported in G studies of the multiple-choice tests for persons, items, and subtests?
3. What are the relative magnitudes of the main-effects VCs reported in G studies for the task/performance ratings tests?
4. What are the relative magnitudes of the VCs for interactions reported in G studies for the task/performance ratings tests?
5. What are the relative magnitudes of the VCs for persons across all G studies?

## METHOD

### Materials

Table 4 summarizes the G studies used in the present paper in terms of the author(s) and date(s) of publication, sample type, academic setting, test from which the data were taken (where relevant), purpose of the test, and type of items used. Notice in Table 4 that the studies cover the time frame from 1982 to the present; that the samples include students learning a variety of different second languages (EFL, ESL, German, & French); that examinees were studying in a number of different settings (university, immigrant, high school, translators, junior college, the Defense Language Institute, and TOEIC) with widely mixed ages; that they were taking a variety of different tests including AACES, COT, ELIPT, SPEAK, TOEFL, TOEIC, TSE, TWE (for full names, see endnote[6]), and a number of local tests; that the tests varied in their purposes (ranging from Engineering-English reading comprehension to pragmatics Role Play Self-assessment); and that there were many different types of items involved (ranging from multiple-choice to self-assessment). Earlier in Table 3, I also showed that the sample sizes ranged widely from 20 to 90,312 and that there were many different G-study designs employed in these investigations.

**Table 4** Summary of General Characteristics of G-Studies Used in the Present Research Synthesis

| Author(s) & Date(s) | Sample | Setting | Test | Purpose | Type of Items |
|---|---|---|---|---|---|
| Brown & Bailey, 1984 | UCLA ESL | University | | Writing | Writing sample |
| Van Weeren & Theunissen, 1987 | US German FL | University | | Pronunciation | Read aloud |
| Van Weeren & Theunissen, 1987 | US French FL | University | | Pronunciation | Read aloud |
| Brown & Ross, 1993 | EFL/ESL | University | TOEFL | Overall English language proficiency | M-C |

| | | | | | |
|---|---|---|---|---|---|
| Brown & Ross, 1993 | EFL/ESL | University | TOEFL | Listening comprehension | M-C |
| Brown & Ross, 1993 | EFL/ESL | University | TOEFL | Grammar | M-C |
| Brown & Ross, 1993 | EFL/ESL | University | TOEFL | Vocabulary and reading comprehension | M-C |
| Brown & Ross, 1993 | EFL/ESL | University | TOEFL | One subtest of vocab and reading comprehension | M-C |
| Lynch & McNamara, 1998 | ESL | Immigrant | AACES | Speaking | Speaking interaction |
| Brown, 1999 | EFL/ESL | University | TOEFL | Overall English language proficiency | M-C |
| Shin, 2002 | Korean EFL | High school | | Reading | M-C |
| Xi, 2003 | ESL | University | SPEAK | Speaking | Tape-mediated listening/speaking |
| Zhang, 2003 | ESL | University | ELIPT | General reading | M-C |
| Zhang, 2003 | ESL | University | ELIPT | Reading comp | M-C |
| Zhang, 2003 | ESL | University | ELIPT | Cloze | M-C |
| Zhang, 2003 | ESL | University | ELIPT | Reading comp | M-C |
| Zhang, 2003 | ESL | University | ELIPT | Vocabulary | M-C |
| Kozaki, 2004 | Japanese EFL | Translators | | Written translations | Writing samples |
| Yoshida, 2004, 2006 | Japanese EFL | Jr. College | | Pronunciation (Dialogue) | Dialogue read aloud |
| Yoshida, 2004, 2006 | Japanese EFL | Jr. College | | Pronunciation (Prose passage) | Prose read aloud |
| Zhang, 2004, 2006 | Japanese & Korean EFL | TOIEC Ages 9 to 90 | TOIEC | Listening & Reading | M-C |
| Zhang, 2004, 2006 | Japanese EFL | TOIEC Ages 10 to 87 | TOIEC | Listening & Reading | M-C |

**Table 4** *(cont.)*

| Author(s) & Date(s) | Sample | Setting | Test | Purpose | Type of Items |
|---|---|---|---|---|---|
| Zhang, 2004, 2006 | Korean EFL | TOIEC Ages 9 to 90 | TOIEC | Listening & Reading | M-C |
| Yamanaka, 2005 | Japanese EFL | High school | | Writing | Writing sample |
| Yamanaka, 2005 | Japanese EFL | High school | | Writing | Writing sample |
| Yamanaka, 2005 | Japanese EFL | High school | | Writing | Writing sample |
| Yamanaka, 2005 | Japanese EFL | High school | | Writing | Writing sample |
| Lee, 2005, 2006 | ESL | University | TSE | Speaking | Tape-mediated listening/speaking |
| Lee & Kantor, 2005, 2007 | ESL | University | TWE | Writing | Writing sample |
| Alharby, 2006 | Saudi EFL | University | | Writing | Writing sample |
| Alharby, 2006 | Saudi EFL | University | | Writing | Writing sample |
| Alharby, 2006 | Saudi EFL | University | | Writing | Writing sample |
| Alharby, 2006 | Saudi EFL | University | | Writing | Writing sample |
| Tang, 2006 | Chinese EFL | High school graduates | COT | Speaking (Form A) | Computerized speaking |
| Tang, 2006 | Chinese EFL | High school graduates | COT | Speaking (Form B) | Computerized speaking |
| Tang, 2006 | Chinese EFL | High school graduates | COT | Speaking (Form C) | Computerized speaking |
| Tang, 2006 | Chinese EFL | High school graduates | COT | Speaking (Form D) | Computerized speaking |
| Banno, 2008 | Chinese JSL | University | | Speaking | Speaking sample |
| Banno, 2008 | Chinese JSL | University | | Speaking | Speaking sample |
| Brown, 2008 | US KFL | University & DLI | | Pragmatics Written Discourse Completion Task | Written shortresponse |
| Brown, 2008 | US KFL | University & DLI | | Pragmatics Oral Discourse Completion Task | Spoken shortresponse |

| | | | | |
|---|---|---|---|---|
| Brown, 2008 | US KFL | University & DLI | Pragmatics Discoures Role Play task | Role-play |
| Brown, 2008 | US KFL | University & DLI | Pragmatics Role Play Self-assessment | Self-assessment |

### Analyses

I began this research synthesis by assembling all available G studies in the field of second and foreign language testing. Once I had selected all the relevant G studies, I further narrowed the focus of the analysis by assigning a common set of labels across all the studies. Then I entered the VCs from the relevant papers into an *Excel* spreadsheet and converted all the VCs into percentages by adding them up within each G study and dividing each VC by the resulting within-study total. Once all VCs were converted to percentages, I sorted through the data according to the design type and facets involved and thus found the patterns that are reported below.

## RESULTS

Researchers in our field (including me) have typically reported the VCs for their G studies with only very brief interpretations and then moved on to the D-study stage which they find more interesting and useful from a practical perspective. In the present research synthesis, I focus on the VCs reported in the G-study stage of the relevant investigations and attempt to determine what their relative magnitudes mean and what the patterns of VCs across studies can tell us.

### What Do VCs Mean?

The first step in interpreting VCs is to recognize that the focus should be on there relative magnitudes. Consider for example the relative magnitudes of the VCs shown in Table 2 above.[7] Notice that the VC for persons (p) is 0.03140424, which is about 13 times larger than the VC of 0.00247421 for subtests (s), almost three times larger than the VC of 0.01098074 for items nested within subtests (i:s), more than four times larger than the VC of 0.00720131 for the ps interaction, and in contrast, only onefifth the size of the VC of 0.16180465 for the pi:s interaction. Put another way, the VC for the pi:s interaction is clearly the largest, with the VC for persons coming in second at about one-fifth the magnitude, followed by much smaller VCs for i:s, ps, and s, in descending order of magnitude. Using such information, testing researchers can interpret the relative importance to the total test score variance of each facet and interaction.

**Table 5** Variance Components from Table 2 and Their Percentages (adapted from Brown & Ross, 1999)

| SOURCE | VC | Percentages |
|---|---|---|
| p | .03140424 | 14.68 |
| s | .00247421 | 1.16 |
| i:s | .01098074 | 5.13 |

| | | |
|---|---|---|
| ps | .00720131 | 3.37 |
| pi:s | .16180465 | 75.66 |
| Total | .21386515 | 100.00 |

These relative magnitudes are easier to interpret and compare (within and across studies) if they are converted to percentages for each VC (relative to the sum of the VCs). Such percentages are shown in the column furthest to the right in Table 5. Examining the percentages in more detail, several patterns emerge. First, the VC for persons indicates the degree to which persons differ from each other. In this case, the relatively large VC for persons (14.68%) indicates that the test is spreading people out to some degree, though this is not a very high percentage of persons variance for such a high-stakes test. Second, VC for subtests indicates the degree to which subtests differ from each other in difficulty. In this case, the relatively small subtests VC (1.16%) indicates that the subtests are of about equal difficulty (i.e., their means do not vary much). Third, the VC for items nested within subtests indicates the degree to which items differ from each other. In this case, the somewhat larger VC for i:s (5.13%) indicates that, to some degree, the items nested within subtests vary more in difficulty than the subtests do, that is, the items nested within subtests VC is more than four times as important as the subtests VC. Fourth, VC for the persons by subtests interaction indicates the degree to which persons differ from each other with regard to which subtests they found difficult or easy. In this case, the VC for the ps interaction (3.37%) indicates that, to some degree, persons' standings (relative to each other) differed across subtests. The VC for the persons by items nested within subtests indicates the degree to which persons differ from each other with regard to which items (nested within subtests) they found difficult or easy. Obviously, the lion's share of variance was due to the VC for the pi:s interaction (75.66%), which shows that, to a large extent, the relative standings of persons differed across items (which are nested within subtests). In other words, relative to each other, different persons answered different items correctly. Since this is also the *highestorder interaction*, it therefore also contains undifferentiated error. In other words, the persons-by-items nested within subtests interaction is confounded with other facets of measurement not explicitly odelled in a particular study design or other unsystematic sources of error (see Shavelson & Webb, 1991, pp. 20-21). Hereafter, I will simply refer to this phenomenon as *undifferentiated error.*

These VCs taken together reveal that the test was spreading examinees (persons) out, but also that, relative to each other, the performances of these persons differed considerably across items (nested within subtests), but to a much lesser degree from subtest to subtest.

## Comparing VCs among Studies

My primary purpose in this research synthesis was not to understand the relative importance of facets and interactions within each study (as discussed in the previous section), but rather to examine all of the studies and understand how any patterns discernable from the entire set of studies can help us understand language tests in general. Sorting the combined results in various ways lead me to examine two types of studies separately: (a) the G studies that were based on multiple-choice tests and

focused on persons, items, and subtests facets (see Table 6) and (b) the G studies that were based on task/performance tests involving raters (see Tables 7a & b).

**Table 6** VCs for M-C Designs with Persons, Items, and Subtests Facets

| Author(s) | Design | Persons | Items | i:s | i:t:s | Subtests | Pi | pi:s | pi:t:s | ps | pt:s | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zhang, 2003 | pi | .0223 | .0271 | | | | .1989 | | | | | .2483 |
| Zhang, 2003 | pi | .0165 | .0316 | | | | .1991 | | | | | .2472 |
| Zhang, 2003 | pi | .0345 | .0090 | | | | .2046 | | | | | .2481 |
| Brown, 1983 | pi:s | .0165 | | .0284 | | .0000 | | .1970 | | .0071 | | .2490 |
| Brown & Ross, 1993 | pi:s | .0314 | | .0110 | | .0025 | | .1618 | | .0072 | | .2139 |
| Brown & Ross, 1993 | pi:s | .0406 | | .0118 | | .0000 | | .1838 | | .0014 | | .2375 |
| Brown & Ross, 1993 | pi:s | .0352 | | .0092 | | .0003 | | .1512 | | .0015 | | .1974 |
| Brown & Ross, 1993 | pi:s | .0361 | | .0119 | | .0006 | | .1561 | | .0033 | | .2080 |
| Brown & Ross, 1993 | pi:s | .0370 | | .0076 | | .0071 | | .1514 | | .0123 | | .2155 |
| Brown, 1999 | pi:s | .0325 | | .0106 | | .0031 | | .1622 | | .0064 | | .2148 |
| Zhang, 2003 | pi:s | .0163 | | .0252 | | .0000 | | .2005 | | .0071 | | .2490 |
| Zhang, 2003 | pi:s | .0182 | | .0186 | | .0046 | | .2011 | | .0020 | | .2445 |
| Zhang, 2004 | pi:s | .0198 | | .0334 | | .0000 | | .1781 | | .0039 | | .2353 |
| Zhang, 2004; 2006 | pi:s | .0200 | | .0355 | | .0000 | | .1766 | | .0034 | | .2355 |
| Zhang, 2004; 2006 | pi:s | .0197 | | .0339 | | .0008 | | .1771 | | .0036 | | .2350 |
| Brown, 2006; 2008 | pi:s | .4659 | | .0473 | | .0000 | | .4927 | | .0111 | | 1.0170 |
| Shin, 2002 | pi:t:s | .0225 | | .0000 | .0257 | .0000 | | .0000 | .1936 | | .0000 | .2418 |

**Table 7a** VCs for G Studies of Task/Performance Tests (Main Effects and Persons Interactions)

| Author(s) | Design | Persons | Items | i:s | Categories | Subtests | Raters | pi | pi:s | pi:c | pc | ps | pr | pir | pi:sr | pi:cr | picr | pcr | pic | psr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alharby, 2006 | pr | .3430 | | | | | .3590 | | | | | | .5050 | | | | | | | |
| Alharby, 2006 | pr | .3350 | | | | | .3170 | | | | | | .4790 | | | | | | | |
| van Weeren & Theunissen, 1987 | pir | .0110 | .0200 | | | | .0050 | .0750 | | | | | .0010 | .0660 | | | | | | |
| van Weeren & Theunissen, 1987 | pir | .0030 | .0280 | | | | .0020 | .0290 | | | | | .0010 | .1100 | | | | | | |
| Lynch & McNamara, 1998 | pir | .5800 | .0020 | | | | .0570 | .0030 | | | | | .0220 | .0030 | | | | | | |
| Lee & Kantor, 2005, 2007 | pir | .5100 | .1400 | | | | .0000 | .2600 | | | | | .0100 | .3800 | | | | | | |
| Lee, 2005, 2006 | pir | .6690 | .0220 | | | | .0000 | .2250 | | | | | .0240 | .3600 | | | | | | |
| Xi, 2003 | pir | 38.9400 | .1400 | | | | .4100 | 2.4200 | | | | | 7.0400 | 12.7500 | | | | | | |
| Tang, 2006 | pir | .9090 | .1500 | | | | .0000 | .5690 | | | | | .2640 | .9560 | | | | | | |
| Tang, 2006 | pir | .7690 | .7840 | | | | .0000 | .3590 | | | | | .2990 | 1.0310 | | | | | | |
| Tang, 2006 | pir | .8560 | .0980 | | | | .0000 | .6420 | | | | | .3150 | .8170 | | | | | | |
| Tang, 2006 | pir | 1.0300 | .0300 | | | | .0000 | .5950 | | | | | .169 | .8590 | | | | | | |
| Banno, 2008 | pir | 1.354 | .0690 | | | | .1350 | .0760 | | | | | .1250 | .2930 | | | | | | |
| Banno, 2008 | pir | 28.2380 | 1.4940 | | | | 2.9230 | 2.0170 | | | | | 3.0690 | 4.5210 | | | | | | |
| Kozaki, 2004 | picr | .1110 | .0590 | | .0000 | | .0710 | .0170 | | | .0080 | | .1110 | .1030 | | | .1760 | .0220 | .0090 | |
| Brown & Bailey, 1984 | pcr | 1.9500 | | | .8000 | | .1600 | | | | .7000 | | 1.1700 | | | | | 1.8100 | | |
| Yamanaka, 2005 | pcr | 1.0100 | | | .0200 | | 6.3600 | | | | .1000 | | 1.3100 | | | | | 1.0100 | | |
| Yamanaka, 2005 | pcr | .7700 | | | .7700 | | 3.3700 | | | | .0500 | | 1.3600 | | | | | .8100 | | |
| Yamanaka, 2005 | pcr | .4300 | | | .2300 | | 6.0000 | | | | .5100 | | .6200 | | | | | 2.0800 | | |
| Yamanaka, 2005 | pcr | .4500 | | | .0000 | | 1.8500 | | | | .8200 | | .7100 | | | | | 1.4200 | | |
| Alharby, 2006 | pcr | .2420 | | | .0060 | | .5510 | | | | .0740 | | .1550 | | | | | .3700 | | |
| Alharby, 2006 | pcr | .2020 | | | .0040 | | .5260 | | | | .0820 | | .1420 | | | | | .3540 | | |
| Brown, 2008 | pi:sr | .3987 | | .0404 | | .0000 | .0527 | | .2048 | | | .0172 | .1135 | | .3285 | | | | | .0104 |
| Brown, 2008 | pi:sr | .3151 | | .0561 | | .0000 | .0170 | | .2712 | | | .0148 | .0682 | | .2987 | | | | | .0000 |
| Brown, 2008 | pi:sr | .7993 | | .0205 | | .0069 | .0227 | | .0894 | | | .0062 | .0741 | | .3071 | | | | | .0062 |
| Brown, 2008 | pi:s | .4659 | | .0473 | | .0000 | | | .4927 | | | .0111 | | | | | | | | |
| Yoshida (2004, 2006) | pi:cr | .0978 | | .0187 | .0004 | | .2995 | | | .0210 | .0086 | | .0825 | | | .1806 | | .0338 | | |
| Yoshida (2004, 2006) | pi:c r | .0932 | | .0339 | .0002 | | .3749 | | | .0346 | .0100 | | .0543 | | | .1843 | | .0396 | | |

**Table 7b** VCs for G Studies of Task/Performance Tests (Main Effects and Non-persons Interactions)

| Author(s) | ri | ri:s | ri:c | rc | rs | ci | rci | Total |
|---|---|---|---|---|---|---|---|---|
| Alharby, 2006 | | | | | | | | 1.2070 |
| Alharby, 2006 | | | | | | | | 1.1310 |
| Van Weeren & Theunissen, 1987 | .0140 | | | | | | | .1920 |
| van Weeren & Theunissen, 1987 | .0340 | | | | | | | .2070 |
| Lynch & McNamara, 1998 | .0000 | | | | | | | .6670 |
| Lee & Kantor, 2005, 2007 | .0000 | | | | | | | 1.3000 |
| Lee, 2005, 2006 | .0030 | | | | | | | 1.3030 |
| Xi, 2003 | .1400 | | | | | | | 61.8400 |
| Tang, 2006 | .0040 | | | | | | | 2.8520 |
| Tang, 2006 | .0110 | | | | | | | 3.2530 |
| Tang, 2006 | .0040 | | | | | | | 2.7320 |
| Tang, 2006 | .0250 | | | | | | | 2.7080 |
| Banno, 2008 | .0370 | | | | | | | 2.0900 |
| Banno, 2008 | .3990 | | | | | | | 42.6610 |
| Kozaki, 2004 | .0410 | | | .0120 | | .0050 | .0210 | .7870 |
| Brown & Bailey, 1984 | | | | .1300 | | | | 6.7200 |
| Yamanaka, 2005 | | | | .6000 | | | | 10.4100 |
| Yamanaka, 2005 | | | | 1.3600 | | | | 8.4900 |
| Yamanaka, 2005 | | | | .0200 | | | | 9.8900 |
| Yamanaka, 2005 | | | | .0300 | | | | 5.2800 |
| Alharby, 2006 | | | | .0150 | | | | 1.4130 |
| Alharby, 2006 | | | | .0160 | | | | 1.3260 |
| Brown, 2008 | | .0154 | | | .0009 | | | 1.1824 |
| Brown, 2008 | | .0158 | | | .0000 | | | 1.0569 |
| Brown, 2008 | | .0223 | | | .0029 | | | 1.3577 |
| Brown, 2008 | | | | | | | | 1.0170 |

| | | | |
|---|---|---|---|
| Yoshida (2004, 2006) | .0309 | .0287 | 0.8026 |
| Yoshida (2004, 2006) | .0498 | .0254 | 0.9002 |

Given that the sum of the VCs is different from study to study, making comparisons across studies was facilitated by converting each VC to a percentage of the total variance in that G study. In this way, the VCs were all put on a scale that would allow for examination across studies of their relative magnitudes within each study. These percentages are shown in Tables 8, 9a, and 9b, which will be discussed in two sections: one for the G studies of multiple-choice tests and another for the G studies of task/ performance ratings tests.

***G Studies of Multiple-Choice Tests***

Notice in Table 8 that the G studies have been rearranged so that the different designs (in the second column) are grouped together and arranged from the very simple pi designs in Zhang (2003) to the more complex pi:t:s design in Shin (2002). The G studies in this section are all based on multiple-choice tests and focus on three facets: *persons* (the column shaded with down-diagonal lines), *items* (the columns shaded in light grey), and *subtests* (the column in the middle of the table with no shading). I will consider each of these facets in turn. Note that the three main-effects facets (i.e., persons, items, and subtests) are not sources of error. Instead they represent the relative degree to which persons, items, or subtests vary. However, the variances for interactions of persons and items as well as of persons and subtests are potential sources of error because they represent how the persons' performances differed, or were *in*consistent, from item to item, or subtest to subtest, or both.

First, the percentages of *persons* variance shown in Table 8 ranged considerably from 6.53% to 17.82%. Indeed, the percentage of persons variance was remarkably low in all of these G studies, considering that the NRT purpose of such tests is to spread the persons out along a continuum of L2 abilities. In addition, in all cases, the percentage of persons variance was much lower than the percentage of persons x items interaction variance (highlighted in darker grey).

Second, the percentages of *items* variance (whether for I, i:s, or i:t:s) ranged from 3.54 to 15.06. It is not surprising that tests designed to spread people out will be built from items that differ in difficulty. In more than half the studies, items produced higher proportions of variance than persons did, but in all cases, items accounted for less variance than the higher order interactions involving persons and items.

Third, the percentages of *subtests* variance are small with only one out of the 13 exceeding 3%. Indeed, in six out of the 13 G studies where subtests were an issue, subtests variance is zero. These results indicate that subtests typically either do not vary in difficulty at all or vary only slightly.

Fourth, the percentages of *persons-by-items* (whether pi, or pi:s, or pi:t:s) interactions variance (shaded in darker grey) were consistently very high relative to the other facets and interactions, ranging from 70.27% to 82.46%. This indicates that the persons-by-items interaction is very large. Since the persons-by-items interaction is the highest order interaction in each G study, it also includes other undifferentiated error. Thus, it is *impossible* to say that these values represent variance due solely to the persons-by-items interaction. Nonetheless, the large relative percentage accounted for

by these interactions and undifferentiated error is clearly trying to tell us something (more about this anon).

Fifth, the variances accounted for by the *persons-by-subtests* interactions (whether ps or pt:s) only exceeded 3% in three out of 13 cases (the ones in bold typeface), which of course, also means that 10 out of 13 were not even that high. Notice that the highest of these three at 5.73% was associated with a percentage of *subtests* variance which is also in boldfaced type, and that all three were associated with at least some subtests variance (i.e., not zero variance). This may indicate that such interactions are associated with subtests that themselves differ at least somewhat in difficulty.

**Table 8** VC Percentages for M-C Designs with Persons, Items, and Subtests Facets

| Author(s) | Design | Persons | Items | i:s | i:t:s | Subtests | pi | pi:s | pi:t:s | ps | pt:s | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zhang, 2003 | pi | **8.99** | **10.90** | | | | **80.11** | | | | | 100.00 |
| Zhang, 2003 | pi | **6.68** | **12.78** | | | | | **80.53** | | | | 100.00 |
| Zhang, 2003 | pi | **13.89** | **3.65** | | | | | **82.46** | | | | 100.00 |
| Brown, 1982, 1984 | pi:s | **6.63** | | **11.41** | | .00 | | | **79.12** | 2.85 | | 100.00 |
| Brown & Ross, 1993 | pi:s | **14.68** | | **5.13** | | 1.16 | | | **75.66** | **3.37** | | 100.00 |
| Brown & Ross, 1993 | pi:s | **17.07** | | **4.96** | | .00 | | | **77.39** | .57 | | 100.00 |
| Brown & Ross, 1993 | pi:s | **17.82** | | **4.68** | | .14 | | | **76.60** | .75 | | 100.00 |
| Brown & Ross, 1993 | pi:s | **17.37** | | **5.72** | | .29 | | | **75.05** | 1.57 | | 100.00 |
| Brown & Ross, 1993 | pi:s | **17.17** | | **3.54** | | **3.30** | | | **70.27** | **5.73** | | 100.00 |
| Brown, 1999 | pi:s | **15.12** | | **4.93** | | 1.42 | | | **75.52** | **3.00** | | 100.00 |
| Zhang, 2003 | pi:s | **6.53** | | **10.12** | | .00 | | | **80.51** | 2.84 | | 100.00 |
| Zhang, 2003 | pi:s | **7.43** | | **7.60** | | 1.88 | | | **82.26** | .83 | | 100.00 |
| Zhang, 2004, 2006 | pi:s | **8.43** | | **14.19** | | .00 | | | **75.71** | 1.67 | | 100.00 |
| Zhang, 2004, 2006 | pi:s | **8.50** | | **15.06** | | .00 | | | **75.00** | 1.45 | | 100.00 |
| Zhang, 2004, 2006 | pi:s | **8.36** | | **14.41** | | .34 | | | **75.36** | 1.52 | | 100.00 |
| Shin, 2002 | pi:t:s | **9.30** | | .00 | **10.62** | .00 | | .00 | **80.07** | | .00 | 100.00 |

*Those values over 3.00 percent are in bold

### *G Studies of Task/Performance Ratings Tests*

I will now turn to those G studies of task/performance tests based on ratings. These studies (shown in Tables 9a & b) all include raters and ten include categories facets. Notice that the percentages of variance for *persons* (shown in the shaded column to the left) are as high as 86.96%. Indeed, persons account for percentages in double digits in 22 out of the 28 studies. As such, these percentages are comparable to, and in a number

of cases higher than, those shown for persons in Table 8 (with the exception of the second one reported in Table 9a from Weeren and Theunissen, 1987). Also salient is the fact that the percentage of variance reported in Lynch and McNamara (1998) is exceptionally high at 86.96%. Thus, in all but one of these tests, at least some of the variance or even very high percentages of variance are due to persons, which is where we would like it to be for norm-referenced tests.

Table 9a also indicates that *items* variance (either I or i:s, the first two columns shaded in light grey) was above 3% in 13 out of the 28 G studies where it was applicable – with those 13 ranging from 3.30% to 25.10%; *categories* variance was over 3% in only two of the 10 cases in which it was applicable, but those two were 9.07% and 11.90%; *subtests* variance (with no shading) was non-existent or nearly non-existent in the four G studies where it was relevant; but *raters* variance (shaded in dark grey) was generally higher, ranging from 4.46% to 61.10%, in 15 out of the 27 G studies where it was applicable.

Turning now to the interaction effects, consider the two-way interactions (shaded in medium grey in Table 9a). Since items and categories interactions seem to be functioning in a similar manner, I will consider them together: percentages of variance ranging from 3.84% to 48.45% were reported for pi, pi:s, or pc interactions in the 21 out of the 27 cases where they were applicable. The percentage of ps interaction variance was zero or close to zero in the four cases where it applied. The percentage of pr interaction variance also ranged from 3.30% to 42.35% in 23 out of 27 G studies where it was applicable.

Next, consider the three-way and four-way interactions beginning with the columns shaded with up-diagonal lines (i.e., pir, pi:sr, pi:cr, picr, & pcr). With the exception of the pir interaction reported in Lynch and McNamara (1998), all of the three-way and four-way interactions that involve both items and raters are sources of at least some error variance ranging from10.60% to 53.14%, with the all but three being higher than 20%. In addition, the seven of the eight pcr interactions (excepting the one that is *not* the highest-order interaction in its design) ranged from 4.21% to 26.93%. I must emphasize again that all of these highest-order interactions also contain undifferentiated error that is not accounted for in their designs.

Another interesting result reported in Table 9a, is that, for one two-way interaction (i.e., ps), and for those three-way and four-way interactions that did not include both items and raters (i.e., ps, pic, and psr in columns with no shading), none contributed to the error variance in even a small way.

**Table 9a** VC Percentages for G Studies of Task/Performance Tests (Main Effects and Persons Interactions)

| Author(s) | Design | persons | items | i:s | categories | subtests | raters | pi | pi:s | pi:c | pc | ps | pr | pir | pi:sr | pi:cr | picr | pcr | pic | psr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Allarby, 2006 | pr | **28.42** | | | | | **29.74** | | | | | | **41.84** | | | | | | | |
| Allarby, 2006 | pr | **29.62** | | | | | **28.03** | | | | | | **42.35** | | | | | | | |
| van Weeren & Theunissen, 1987 | pir | **5.73** | **10.42** | | | | 2.60 | **39.06** | | | | | .52 | **34.38** | | | | | | |
| van Weeren & Theunissen, 1987 | pir | 1.45 | **13.53** | | | | .97 | **14.01** | | | | | .48 | **53.14** | | | | | | |
| Lynch & McNamara, 1998 | pir | **86.96** | .30 | | | | **8.55** | .45 | | | | | **3.30** | .45 | | | | | | |
| Lee & Kantor, 2005, 2007 | pir | **39.23** | **10.77** | | | | .00 | **20.00** | | | | | .77 | **29.23** | | | | | | |
| Lee, 2005, 2006 | pir | **51.34** | 1.69 | | | | .00 | **17.27** | | | | | 1.84 | **27.63** | | | | | | |
| Xi, 2003 | pir | **62.97** | .23 | | | | .66 | **3.91** | | | | | **11.38** | **20.62** | | | | | | |
| Tang, 2006 | pir | **31.87** | **5.26** | | | | .00 | **19.95** | | | | | **9.26** | **33.52** | | | | | | |
| Tang, 2006 | pir | **23.64** | **25.10** | | | | .00 | **11.04** | | | | | **9.19** | **31.69** | | | | | | |
| Tang, 2006 | pir | **31.33** | **3.59** | | | | .00 | **23.50** | | | | | **11.53** | **29.90** | | | | | | |
| Tang, 2006 | pir | **38.04** | 1.11 | | | | .00 | **21.97** | | | | | **6.24** | **31.72** | | | | | | |
| Banno, 2008 | pir | **64.78** | **3.30** | | | | **6.46** | **3.64** | | | | | **5.98** | **14.02** | | | | | | |
| Banno, 2008 | pir | **66.19** | **3.50** | | | | **6.84** | **4.73** | | | | | **7.19** | **10.60** | | | | | | |
| Kozaki, 2004 | picr | **14.10** | **7.50** | | .00 | | **9.02** | 2.16 | | | 1.02 | | **14.10** | **13.09** | | | **22.36** | 2.80 | 1.14 | |
| Brown & Bailey, 1984 | pcr | **29.02** | | | **11.90** | | 2.38 | | | | **10.42** | | **17.41** | | | | | **26.93** | | |
| Yamanaka, 2005 | pcr | **9.70** | | | .19 | | **61.10** | | | | .96 | | **12.58** | | | | | **9.70** | | |
| Yamanaka, 2005 | pcr | **9.07** | | | **9.07** | | **39.69** | | | | .59 | | **16.02** | | | | | **9.54** | | |
| Yamanaka, 2005 | pcr | **4.35** | | | 2.33 | | **60.67** | | | | **5.16** | | **6.27** | | | | | **21.03** | | |
| Yamanaka, 2005 | pcr | **8.52** | | | .00 | | **35.04** | | | | **15.53** | | **13.45** | | | | | **26.89** | | |
| Allarby, 2006 | pcr | **17.13** | | | .42 | | **39.00** | | | | **5.24** | | **10.97** | | | | | | | |
| Allarby, 2006 | pcr | **15.23** | | | .30 | | **39.67** | | | | **6.18** | | **10.71** | | | | | | | |
| Brown, 2008 | pi:sr | **33.72** | | **3.42** | | .00 | **4.46** | | **17.32** | | | 1.45 | **9.60** | | **27.78** | | | | | .88 |
| Brown, 2008 | pi:sr | **29.81** | | **5.31** | | .00 | 1.61 | | **25.66** | | | 1.40 | **6.46** | | **28.26** | | | | | .00 |
| Brown, 2008 | pi:sr | **58.87** | | 1.51 | | .51 | 1.67 | | **6.59** | | | .46 | **5.46** | | **22.62** | | | | | .45 |
| Brown, 2008 | pi:s | **45.81** | | **4.65** | | .00 | | | **48.45** | | | 1.09 | | | | | | | | |
| Yoshida (2004, 2006) | pi:cr | **12.19** | | 2.34 | 0.05 | | **37.32** | | | 1.07 | 2.61 | | **10.28** | | | **22.50** | | **4.21** | | |
| Yoshida (2004, 2006) | pi:cr | **10.35** | | **3.76** | 0.02 | | **41.65** | | | 1.11 | **3.84** | | **6.03** | | | **20.48** | | **4.40** | | |

*Those values over 3.00 percent are in bold.

**Table 9(b)** VC Percentages for G Studies of Task/Performance Tests (Non-persons Interactions)

| Author(s) | pi | pi:s | pi:c | pc | ps | pr | pir | pi:sr | pi:cr | picr | pcr | pic | psr | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alharby, 2006 | | | | | | | | | | | | | | 100.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Alharby, 2006 | | | | | | 100.00 |
| van Weeren & Theunissen, 1987 | **7.29** | | | | | 100.00 |
| van Weeren & Theunissen, 1987 | **16.43** | | | | | 100.00 |
| Lynch & McNamara, 1998 | .00 | | | | | 100.00 |
| Lee & Kantor, 2005, 2007 | .00 | | | | | 100.00 |
| Lee, 2005, 2006 | .23 | | | | | 100.00 |
| Xi, 2003 | .23 | | | | | 100.00 |
| Tang, 2006 | .14 | | | | | 100.00 |
| Tang, 2006 | .34 | | | | | 100.00 |
| Tang, 2006 | .15 | | | | | 100.00 |
| Tang, 2006 | .92 | | | | | 100.00 |
| Banno, 2008 | 1.77 | | | | | 100.00 |
| Banno, 2008 | .94 | | | | | 100.00 |
| Kozaki, 2004 | **5.21** | | 1.52 | .64 | 2.67 | 100.00 |
| Brown & Bailey, 1984 | | | 1.93 | | | 100.00 |
| Yamanaka, 2005 | | | **5.76** | | | 100.00 |
| Yamanaka, 2005 | | | **16.02** | | | 100.00 |
| Yamanaka, 2005 | | | .20 | | | 100.00 |
| Yamanaka, 2005 | | | .57 | | | 100.00 |
| Alharby, 2006 | | | 1.06 | | | 100.00 |
| Alharby, 2006 | | | 1.21 | | | 100.00 |
| Brown, 2008 | | 1.30 | | .07 | | 100.00 |
| Brown, 2008 | | 1.49 | | .00 | | 100.00 |
| Brown, 2008 | | 1.64 | | .22 | | 100.00 |

| | | | |
|---|---|---|---|
| Brown, 2008 | | | 100.00 |
| Yoshida (2004, 2006) | **3.85** | **3.58** | 100.00 |
| Yoshida (2004, 2006) | **5.53** | **2.82** | 100.00 |

Table 9b shows the ri, ri:s, ri:c, rc, rs, ci, and rci interactions. These have been put in a separate table and were not mentioned earlier because the focus of this paper is on norm-referenced testing. In G theory applications to norm-referenced testing, only the persons VC and the VCs for interactions with persons (i.e., those shown in Table 9a) play a role in calculating generalizability coefficients (G coefficients). Nonetheless, these non-persons interactions play a role in absolute decisions, or criterion-referenced testing, so I briefly present them in Table 9b, which shows that only nine of the 33 nonpersons interactions are above 3%. The remaining 24 non-persons interactions have very low percentages of variance and would be of little interest even if these tests had been designed for criterion-referenced tests purposes.

## What Do These G Studies Combined Tell us About Norm-referenced Language Testing?

Table 10 shows the percentages of persons variance accounted for in the G studies included in this research synthesis. Notice that the five columns to the left show the G studies that were scaled dichotomously (all but two were multiple-choice; the exceptions were in van Weeren & Theunissen, 1987, where raters scored whether specific phonological features were present or absent), while the five columns to the right show the studies that were polytomously scaled (all were task/performance ratings). The fifth and sixth columns down the middle of the table show the percentages of persons variance (in bold-faced type) accounted for in each of the studies. These persons variance percentages are sorted from low to high for both types of tests, and they are placed in contiguous columns for easy comparison. I focus on persons variance in this table because of its importance in calculating generalizability coefficients for norm-referenced tests. Generally speaking, norm-referenced tests function more dependably if there are relatively high proportions of persons variance.

**Table 10** Persons Variance Percentages for Designs Based on Dichotomous Scales (Mostly M-C) and Polytomous Scales (Task/Performance Ratings)

| Dichotomous-Scale Studies | Design | N-size | SL/ FL?* | Persons % VC | Persons % VC | Polytomous-Scale Studies | Design | N-size | SL/ FL?* |
|---|---|---|---|---|---|---|---|---|---|
| van Weevren & Theunissen, 1987 | pir | 26 | GFL | **1.45** | | | | | |
| | | | | | **4.35** | Yamanaka, 2005 | pcr | 20 | EFL |
| van Weeren & Theunissen, 1987 | pir | 29 | FFL | **5.73** | | | | | |
| Zhang, 2003 | pi:s | 94 | ESL | **6.53** | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Brown, 1982, 1984 | pi:s | 78 | ESL | **6.63** | | | | | |
| Zhang, 2003 | pi | 94 | ESL | **6.68** | | | | | |
| Zhang, 2003 | pi:s | 94 | ESL | **7.43** | | | | | |
| Zhang, 2004, 2006 | pi:s | 45,156 | EFL | **8.36** | | | | | |
| Zhang, 2004, 2006 | pi:s | 90,312 | EFL | **8.43** | | | | | |
| Zhang, 2004, 2006 | pi:s | 45,156 | EFL | **8.50** | | | | | |
| | | | | | **8.52** | Yamanaka, 2005 | pcr | 20 | EFL |
| Zhang, 2003 | pi | 94 | ESL | **8.99** | | | | | |
| | | | | | **9.07** | Yamanaka, 2005 | pcr | 20 | EFL |
| Shin, 2002 | pi:t:s | 157 | EFL | **9.30** | | | | | |
| | | | | | **9.70** | Yamanaka, 2005 | pcr | 20 | EFL |
| | | | | | **10.35** | Yoshida (2004, 2006) | pi:cr | 60 | EFL |
| | | | | | **12.19** | Yoshida (2004, 2006) | pi:cr | 60 | EFL |

**Table 10** *(cont.)*

| **Dichotomous-Scale Studies** | **Design** | **N-size** | **SL/ FL?*** | **Persons % VC** | **Persons % VC** | **Polytomous-Scale Studies** | **Design** | **N-size** | **SL/ FL?*** |
|---|---|---|---|---|---|---|---|---|---|
| Zhang, 2003 | pi | 94 | ESL | **13.89** | | | | | |
| | | | | | **14.10** | Kozaki, 2004 | picr | 20 | EFL |
| Brown & Ross, 1993 | pi:s | 20,000 | ESL | **14.68** | | | | | |
| Brown, 1999 | pi:s | 15,000 | ESL | **15.12** | | | | | |
| | | | | | **15.23** | Alharby, 2006 | pcr | 233 | EFL |
| Brown & Ross, 1993 | pi:s | 20,000 | ESL | **17.07** | | | | | |
| | | | | | **17.13** | Alharby, 2006 | pcr | 233 | EFL |
| Brown & Ross, 1993 | pi:s | 20,000 | ESL | **17.17** | | | | | |
| Brown & Ross, 1993 | pi:s | 20,000 | ESL | **17.37** | | | | | |
| Brown & Ross, 1993 | pi:s | 20,000 | ESL | **17.82** | | | | | |
| | | | | | **23.64** | Tang, 2006 | pir | 6818 | EFL |
| | | | | | **28.42** | Alharby, 2006 | pr | 233 | EFL |
| | | | | | **29.02** | Brown & Bailey, 1984 | pcr | 50 | ESL |
| | | | | | **29.62** | Alharby, 2006 | pr | 233 | EFL |
| | | | | | **29.81** | Brown, 2008 | pi:sr | 53 | KFL |
| | | | | | **31.33** | Tang, 2006 | pir | 3243 | EFL |

| | | | | |
|---|---|---|---|---|
| **33.72** | Brown, 2008 | pi:sr | 53 | KFL |
| **34.87** | Tang, 2006 | pir | 9351 | EFL |
| **38.04** | Tang, 2006 | pir | 1099 | EFL |
| **39.23** | Lee & Kantor, 2005, 2007 | pir | 488 | ESL |
| **45.81** | Brown, 2008 | pi:s | 53 | KFL |
| **51.34** | Lee, 2005, 2006 | pir | 261 | ESL |
| **58.87** | Brown, 2008 | pi:sr | 53 | ESL |
| **62.97** | Xi, 2003 | pir | 20 | ESL |
| **64.78** | Banno, 2008 | pir | 6 | JFL |
| **66.19** | Banno, 2008 | pir | 6 | JFL |
| **86.96** | Lynch & McNamara, 1998 | pir | 83 | ESL |

*GFL = German as a Foreign Language; FFL = French as a Foreign Language; KFL = Korean as a Foreign Language; JFL = Japanese as a Second Language; ESL = English as a Second Language; EFL = English as a Foreign Language

Notice that, overall, the polytomously scaled studies based on task/performance ratings (which are on scales from 1-4, 1-5, 1-6, 1-7, or 1-20) produced higher percentages of persons variance than the dichotomously scaled studies (on right-wrong 0-1, or present or absent scales 1-0 scales). In more detail, it seems clear that the studies with the highest percentages of persons variance are those that were polytomously scaled. Put another way, almost two-thirds of the studies (17 out of 26) involving polytomous ratings have percentages of variance higher than 23.64, which is higher than all 18 of the other G studies dichotomously scaled tests. Looked at a different way, 18 of the 27 studies with the lowest percentage of variance are based on dichotomously scaled tests. In general, then, one overall pattern emerges here (with a few exceptions that will be discussed below): polytomously scaled task/performance tests tend to produce higher percentages of persons variance than dichotomously scaled (mostly multiple-choice) tests do.

## DISCUSSION

In this section, I will provide direct answers to the research questions posed above. To that end, I will use those research questions as sub-headings.

### 1. What are the relative magnitudes of the main-effects VCs reported in G studies of the multiple-choice tests for persons, items, and subtests?

The VC percentages for main-effects facets in the G studies of the multiple-choice tests for persons, items, and subtests (shown in Tables 8, 9a, & 9b) tell us that:

a. The *persons* variances (whether for p) accounted for in these studies were consistently detectable and ranged from small to fairly low percentages.[8] This means that persons differed from each other in terms of the abilities being tested. However, the persons variances were generally low for norm-referenced purposes.
b. *Items* variances (whether for i, i:s, or i:t:s) ranged from small to fairly low, which indicates that items differed somewhat from each other in difficulty. This is not particularly surprising given that norm-referenced tests are usually built from items that differ considerably in difficulty from .30 to .70 (in classical theory approaches), or -3.00 to +3.00 logits (in item-response theory approaches).
c. The *subtests* variances were largely very small in these G studies, which means that subtests differed very little from each other in difficulty.

**2. What are the relative magnitudes of the VCs for interactions reported in G studies of the multiple-choice tests for persons, items, and subtests?**

The relative percentages for the VCs reported in studies of the multiple-choice tests for interactions of persons, items, and subtests (shown in Tables 8 & 9) indicate that:

a. The *persons-by-items* interaction variances (whether for pi, pi:s, or pi:t:s) were very high when compared to the main-effects facets. Persons-by-items interactions indicate the degree to which persons differed with regard to which items they found easy or difficult. Thus students who got the same scores may have gotten those scores by answering different items correctly. Persons-byitems interactions have important implications in terms of the degree to which we can make claims about any hierarchical structure for our norm-referenced language tests. Remember, however, that the interaction is the highest order interaction in these G studies, and as such, it also includes other undifferentiated error.
b. Less than one-fourth of the *persons-by-subtests* interactions (whether ps or pt:s) were even small. Thus persons performances varied little relative to each other across subtests. Note that all of these small interactions were associated with at least some variance for the subtests themselves, which may indicate that such rare interactions, when they occur, tend to be associated with the few subtests that themselves differ in difficulty.

**3. What are the relative magnitudes of the main-effects VCs reported in G studies for the task/performance ratings tests?**

The relative percentages of the main-effects variances reported for the task/ performance ratings tests generally indicate that:

a. The *persons* variances in all but one of the G studies shown in Table 9a account for small to very high percentages of variance, and in most cases, they are relatively high compared to three of the other main-effects variances (items,

categories, and subtests). Thus large proportions of the variance in these tests are focused on the persons, meaning that the persons vary in their scores, which is where the focus should be in norm-referenced tests.

b. As for *items* variance, in those studies that had items, about half had percentages of items variance that were small to fairly low. This is not problematic because it simply means that sometimes the items varied in difficulty.
c. *Categories* variance was fairly low in two cases out of ten G studies, and *subtests* variance made virtually no contribution to the test variance in any case. These results mean that, for the most part, the categories and subtests did not vary in difficulty.
d. *Raters* variance was generally higher, that is, in over half the G studies, the percentage of raters variance ranged from small to high. This means that, in some studies, raters differed in the severity or leniency of their scores. Such results for raters may have implications for rater training if the goal of the testing policy is to have raters who give roughly equivalent scores. Such a policy is not always necessary (see McNamara, 1996).

**4. What are the relative magnitudes of the VCs for interactions reported in G studies for the task/performance ratings tests?**

For the task/performance ratings tests shown in Table 9a, the relative percentage of variance reported for two-way interactions in G studies indicate:

a. About half of the percentages of variance for the pi, pi:s, and pc interactions for those G studies where they were applicable ranged from small to moderately high. This indicates that, in those G studies, the persons performances differed at least somewhat from each other in terms of how difficult they found items or categories to be.
b. The percentage of ps interaction variance was small in all the studies where it applied. Thus, persons did not appear to differ much in terms of how difficult they found subtests.
c. The percentage of pr interaction variance also ranged from small to moderately high in about 85% of the G studies where it applied. This means that raters often differed in the severity or leniency of their ratings for various examinees more often than not.

The relative percentage of variance reported for three-way and four-way interactions in G studies indicate:

i. All but one of the three-way and four-way persons interactions that involve both items and raters (i.e., pir, pi:sr, pi:cr, & picr) were contributing at least some error variance ranging from fairly low to high percentages. This means that persons generally performed differently from each other on various combinations of items, categories, and raters, that is, they scored high or low in non-systematic patterns across the levels of these facets. ii. The pcr interactions (except for the one that is the highest-order interaction in its design) all contributed small to moderately high percentages. This indicates that some persons performed differently from each other on various combinations of

categories and raters, that is, they scored high or low in non-systematic patterns across the levels of these facets.

iii. The ri, ri:s, ri:c, rc, rs, ci, and rci, interactions are of little interest in normreferenced testing because, in G theory applications to norm-referenced testing, G coefficients are calculated solely on the basis of the persons VC and lower-case delta error, which is based on the VCs for interactions with persons.

**5. What are the relative magnitudes of the VCs for persons across all G studies?**

a. One overall pattern that emerges from the relative percentages of the VCs for persons across all the G studies in this research synthesis is that polytomously scaled task/performance tests tend to produce higher percentages of persons variance than dichotomously scaled (mostly multiple-choice) tests.
b. Another pattern is that those studies that produced the highest amounts of persons variance for each of the types of scoring tend to have been *second* language studies where variance may have been more restricted, and conversely, those that produced the lowest amounts of persons variance for each scoring type tended to generally be *foreign* language studies.

**Putting These Results in Perspective**

Language testers have long recognized that a one item test of any sort is probably not a good idea, so multiple-choice norm-referenced tests typically have 20, 30, 40, 50, or even 60 items that all have high item discrimination values. The results of the present paper reinforce the idea that multiple observations are an important part of good testing practices. In the last few decades, we have tended to focus on language tests that use combinations of items, raters, categories, subtests, etc. This means that language testers find themselves using multiple numbers of such facets. The beauty of G theory is that it allows for D studies to find the right balance of numbers for each facet, while taking into consideration all practical and logistical realities.

What this means with reference to the G studies examined in this paper, is that the impacts of interaction VCs that involve persons (which are the error components) can be minimized by increasing the numbers of instances within whatever facets are involved. For example, in many cases above, the persons-by-items variance was relatively large. To minimize the effects of that source of error, relatively large numbers of items should be used. D studies in each case will help testers determine how many items is enough in each case. Similarly, where persons-by-items and persons-by-raters interactions produced relatively large VCs, the effects of those sources of error can be simultaneously minimized by using relatively large numbers of both items and raters. D studies in such cases will help testers determine what the best trade off is in terms of numbers of items and raters, while taking into account the constraints of the particular testing situations involved.

## CONCLUSIONS

Before closing, I would like to step back a bit and compare the results from G study designs of multiple-choice tests (shown in Table 8) with those for the more G studies of tasks/performances (shown in Table 9a & b).

### The G Studies of Multiple-Choice Tests

What can we learn from the G studies of multiple-choice tests with persons, items, and subtests designs? The fact that *persons* variance does not appear to be as high in many cases as the variance produced by items and the items-by-persons interactions is not surprising. Nonetheless, we should continue to do what we can to enhance persons variance through piloting, item analysis, sensible sampling procedures, and good test administration practices.

In terms of the persons part of the very large persons-by-items interactions found in many of these G studies of multiple-choice tests, we need to recognize that examinees with the same scores are answering different items correctly, or put another way, different students with the same scores may know different aspects of the language, especially when those aspects are narrowly defined discrete-point items scored on a right-wrong, or 0-1, scale. Such persons-by-items interactions may be caused by differences in national educational systems, language curricula, textbooks, teachers, budgets, learning conditions, administrative policies, etc. Such interactions may also result from individual differences in learning preferences, study styles, motivation, personality, anxiety levels, attitudes toward language learning, etc.

With regard to the items part of the very large persons-by-items interactions found systematically throughout the G-studies of multiple-choice tests, we need to recognize that we are often: (a) unjustified in claiming that our language testing scales are hierarchical, (b) unable to clearly describe all of the sources of error variance in such tests, and indeed, (c) unable to state with confidence just what such tests are measuring (beyond effectively spreading the examinees out along a continuum that is somehow related to language learning). However, on a more positive note, the G studies tell us that subtests variance is not a particularly important issue and that we should continue to focus our effort on using ample numbers of items to help minimize the impact on test consistency of the persons-by-items interaction as a source of error.

From the perspective of what we can continue to learn from G theory, it seems clear that pi or pi:s designs have been consistent in their results and have not been very effective at helping us to differentiate sources of error variance. For example, it is not very helpful to know that error is due to persons-by-items interactions (and other undifferentiated error variance). Theoretically speaking, we probably should not bother doing more of these pi or pi:s G studies; they no longer tell us much beyond what we could learn from using the Spearman-Brown formula. Instead, we should design more complex G studies of the same sorts of tests that can examine the impact of persons factors (like differences in national educational systems, language curricula, textbooks, teachers, budgets, learning conditions, administrative policies, learning preferences, study styles, motivation, personality, anxiety levels, attitudes toward language learning, etc.) and items factors (like differences in item content, item type, item format, item difficulty, etc.)

## The G Studies of Task/Performance Ratings Tests

What can we learn from the G studies of task/performance ratings tests included here? In many of these studies, the persons variances accounted for a larger proportion of the test variance than the other main effects of items, categories, and raters. Persons variance is *desirable* variance in a norm-referenced test in the sense that it indicates the degree to which the persons are being spread out. In contrast, items, categories, and subtests variances only show the degree to which there are differences in the difficulty of the levels of these facets. However, raters variances need to be interpreted differently because they indicate the degree to which raters were differing in the severity or leniency of their scores. Such differences have potential test design and testing policy implications.

One of the clearest patterns that surfaced in this research synthesis is that the multiple-choice G studies are not accounting for important sources of error variance while the task/performance ratings designs show a picture of higher proportions of persons variance. Hence, these task/performance ratings tests might rightfully be considered more effective norm-referenced tests than the multiple-choice tests. In addition, the G studies of task/performance ratings tests show more evenly spread error variances due to the various interactions involving persons and are therefore better able to account for the error on these tests.

Put another way, where the simpler pi and pi:s designs showed large amounts of error variance due to persons-by-items interactions or other undifferentiated error, the designs based on tasks/performance ratings tests reveal error variances that are more evenly spread across persons-by-items, persons-by-categories, persons-by-raters, and some of the other higher order interactions (especially ones including both items and raters). More precisely, the highest order interactions (where undifferentiated error is located) in each of the designs shown in Table 9a are relatively small, especially when compared to the highest order interactions (i.e., the pi and pi:s interactions) reported in the G studies of multiple-choice tests shown in Table 8. Consequently, the designs based on task/performance ratings tests probably have much less undifferentiated error variance, are better accounting for the sources of error, and give us a better understanding of the error variances in our tests.

Practically speaking, the designs based on task/performance ratings tests show us that there is much we can do to minimize error variance due to such interactions. We need not rely solely on increasing the number of items to reduce error, but can also consider increasing the numbers of categories, raters, etc. to minimize the impact on generalizability of their interactions with persons.

## Persons VCs Across All the G studies

Another pattern worth considering is that the polytomously scaled task/performance ratings tend to produce higher percentages of persons variance than the dichomtomously scored (mostly multiple-choice) designs. This pattern may have arisen for at least four reasons (all of which should probably be examined and compared in future research).

One possible explanation for the pattern shown in Table 10 may be differences in sample size. The persons sample sizes do vary enormously from 6 to 90,312. However,

there does not appear to be any discernable pattern that would indicate a relationship between sample size and the percentages of persons variance accounted for. Indeed, small and large sample sizes appear throughtout Table 10, with the smallest *n*-size being reported for the two Banno (2008) studies (where $n = 6$) producing the second and third highest percentages of persons variance (64.78% & 66.19%) and the largest sample size (Zhang, 2004, 2006; $n = 90{,}312$) producing among the lowest percentages of persons variance (8.43%). Though it is likely, in my experience, that the relative sample sizes are related to the standard errors of the VCs in the various studies, many researchers (including me) have chosen not to report the standard errors of the VCs in their studies, so there is no way to verify what the effects of sample size might be.

A second, more plausible explanation for the pattern of persons VCs in Table 10 is that the different scoring scales (binary and polytomous scaling) affect the relative magnitude of the persons VCs. Most of the multiple-choice tests use a binary scoring scheme (right/wrong), while task/performance tests tend to use polytomous scaling schemes (e.g., 0-5, 0-6, even 0-20). The polytomous scaling systems may simply make it possible to spread persons out more effectively and discriminate among them at the item level, category level, and especially at the total test score level.

A third potential explanation for the patterns in Table 10 is that larger percentages of persons variance can be accounted for in polytomously scaled task/performance tests with more complex G-study designs than in dichotomously scaled multiplechoice tests simply because additional measurement facets like items, categories, raters, and so forth can more easily be modeled in such designs and because modeling them enables us to account for additional sources of error.

A fourth possible explanation has to do with the nature of dichotomously scaled multiple-choice tests and polytomously scaled task/performance test rating scales. Because the abilities being tested in dichotomously scaled multiple-choice tests are narrowly focused, it may be more likely that persons will differ from one another in what they know (i.e., some students will know some things and other students will know other things because they studied in different language programs, had different teachers, used textbooks, etc.)—thus relatively strong persons by items interactions are more likely and the possibility of persons variance is smaller in narrowly focused tests. In contrast, in polytomously scaled task/performance ratings tests, the raters are giving scores that are based on differences among examinees that are relatively global (even when they are scoring analytically), which could lead in turn to relatively less variance due to interactions of persons with categories, raters, etc., and hence, to the possibility of more persons variance.

One issue that arises is why the Yamanaka (2005), Yoshida (2004, 2006), Kozaki (2004), and Alharby (2006) G studies clearly do not fit what appears to be the general pattern. I first thought that perhaps these studies produced less persons variance overall because the data are from foreign language students rather than second language ones (and the ranges of ability were therefore restricted)—though there are other foreign language studies (e.g., Brown, 2008) that have relatively high percentages of persons variance.

Alternatively, it is interesting to note that these four studies (along with Brown & Bailey, 1984) all included *categories* as a facet (all in pcr designs). Perhaps when raters use categories, they tend to make more narrowly focused judgments and therefore make it more likely that persons will differ from one another in what they know (i.e., some

students will know some categories and other students will know others because they studied in different language programs, had different teachers, used textbooks, etc.)—thus relatively strong pc or pcr interactions are more likely and the possibility of persons variance is smaller.

## Limitations

One potential limitation of this research synthesis is that a number of the G studies reported here are not independent of each other. That is, some sets of G studies come from research reports wherein they are based on the same data.

However, the primary limitation of this research synthesis is that it is not a metaanalysis. This is true because the study does not *statistically analyze* the patterns of variance components. In order for any such meta-analysis statistics to be applied, at very least, standard errors of the variance component estimates would have been necessary, and these have only been reported in a few studies to date. Perhaps future G studies should make a point of reporting the standard errors of the variance component estimates.

Instead of doing meta-analysis, this research synthesis examined the patterns of relative strength for variance components across studies. While this is generally a weaker approach, it is in keeping with the way variance components have typically been interpreted and reported in all fields over the years since Cronbach, Rajaratnam, and Gleser (1963) first proposed the whole notion of generalizability theory. So the implications of this study are probably worth heeding.

Another limitation has to do with the distinction between random and fixed effects discussed earlier in this paper. Given that the variations in researcher's choice of random or fixed facets were many and that a number of researchers chose not to even mention this issue, I felt I had no choice but to ignore this issue. In any case, variations in relative VC strength due to whether facets were defined as fixed or random do not appear to have created additional variation that masked the relatively strong patterns observed in this research synthesis. However, this is an area that would benefit from additional research.

A final limitation is the fact that this research synthesis focused solely on G studies of relative decisions and norm-referenced tests (where only persons variance and interactions involving persons are included). Future research syntheses might benefit from examining G studies of absolute decisions and criterion-referenced tests because those sorts of tests are interesting too, but also because additional non-persons interactions can be studied as sources of error.

## Implications

Typically, testers count on using large numbers of items in multiple-choice tests to minimize the sources of error that are not accounted for. And that works – to a degree. Instead, we should perhaps be designing all of our G studies (and the tests they are based on) to be more encompassing in terms of differentiating types of error. For example, perhaps we should design all of our language tests a bit more carefully to balance factors like numbers of items, subtests, item types, item contents, language points being tested, test methods, reading topics, etc. so we can use G theory to examine

them all at the same time as sources of error variance. If we were to do so, we might come to understand which sources of error are, and are not, relatively important. This approach would be much more fruitful than repeatedly taking the easy path that seems to almost inevitably lead to a large persons-by-items interactions and large proportions of undifferentiated error.

Practically speaking, while we would never believe the results of a test that had only one multiple-choice item, we are often tempted in our tests to use one writing sample, one rater, one rating category, etc.—a temptation we might want to resist. However, we must resist rationally. The G studies of task/performance ratings tests in this research synthesis indicate that interactions of persons with items and raters tend to be relatively more important as sources of error than persons interactions with categories or subtests. Rationally speaking, then, we should probably focus on increasing the numbers of items and raters rather than on increasing the numbers of categories and subtests.

However, there are still testing facets that have not been investigated sufficiently, and the results of the G studies of task/performance ratings tests in this research synthesis tell us that the sources of error differ somewhat from study to study. We should therefore study additional facets as sources of error and do so in each individual language testing situation in order to minimize such facets as sources of error by tailoring our tests to the practical constraints and conditions found in each context. Naturally, such tailoring will involve deciding rationally which facets are worth investing with multiple observations.

I have shown here that G-theory allows language testers and researchers to select any facets that they think might be sources of measurement error and study their relative effects on test variance. I have also shown how the various interaction VCs contribute differentially to the error variance on various sorts of tests. Naturally, such information is useful in its own right, and indeed, we should probably pay much more attention to interpreting the VCs in our G studies. That does not mean we should stop doing D studies. After all, they are useful for calculating G coefficients for various existing and potential combinations of numbers of items, categories, raters, etc. and for providing valuable *what-if* information that can further inform test design and revision decisions.

## ENDNOTES

1 The paper is a revised version of the keynote speech I presented in 2007 at the 10th Annual Academic Forum on English Language Testing in Asia (AFELTA) in Tokyo, Japan.

2 Note that Brennan (1983) provides exact equations for a number of common designs in his Appendix B (pp. 129-132).

3 Notice that I kept the *MS* values at eight places to the right of the decimal point. Such precision is critical in calculating VCs because they are very small values that will therefore be greatly affected by early rounding. In G theory, rounding must be left to the very last step in each process. Also, for more on the logic of these variance component computations, see Brennan (1983, pp. 5, 129-132).

4 For those interested, several G studies have also been conducted in first language testing of writing (Brown, 1988, 1989, 1990a, 1991, 2007; Lane & Sabers, 1989; and Sudweeks, Reeve, & Bradshaw, 2005) and in testing linguistic minorities (Solano-Flores & Li, 2006).

5 In recent years, the distinction between *raters* and *ratings* has become a topic of interest to some G theory researchers (for a discussion, see Tang, 2006, pp. 38-39; Lee & Kantor, 2005, pp. 5-7, 37-38).

In short, *raters* involves having one rater in each level of the facet, while *ratings* may have different raters within each level of the facet (based on the belief that these raters are drawn from a fairly homogeneous universe of all possible raters). Lee and Kantor (2005, p. 20) compared G study results for both definitions and found very similar (near zero) estimates of the proportion of rater-related variance, and pr interaction variances (though the *raters* design did have a persons VC of .71 compared to the *ratings* design persons VC of .54, and there was less pir interaction variance including undifferentiated error in the *raters* design with .30 than in the *ratings* design with .40). All in all, this issue probably did not have any great effect on the results of this research synthesis, and in any case, it was only an issue in three of the papers used in this research synthesis: Lee (2005, 2006), Lee and Kantor (2005, 2007), and Tang (2006).

6 AACES=Australian Assessment of Communicative English Skills, COT=Computerized Oral Test, ELIPT=English Language Institute Placement Test, SPEAK=Speaking Proficiency English Assessment Kit, TOEFL=Test of English as a Foreign Language, TOEIC=Test of English for International Communication, TSE=Test of Spoken English, TWE=Test of Written English.

7 In interpreting these results, remember that items are nested within subtests, and therefore, as Shavelson and Webb (1991, p. 74) put it: "it is not meaningful to think about the effect *i:s* as the confounding of separate effects *i* and *is*. No separate *i* effect exists because an item score cannot be interpreted independently of the scale it is in. And no separate *is* interaction effect exists because one cannot think of items having different relative standings across scales [i.e., subtests] – items belong to only one scale."

8 Note: In order to avoid using numbers in this Discussion section, I will consistently use the following terminology to indicate certain percentage ranges: Small = 3% to 9.99%; Fairly low = 10% to 24.99; Moderately high = 25% to 49.99%; High = 50% to 69.99%; and Very high = 70% to 100%. The actual percentages can be found in the **RESULTS** section.

## REFERENCES

Abeywickrama, P. S. (2007). *Measuring the knowledge of textual cohesion and coherence in learners of English as a second language (ESL).* (Unpublished PhD dissertation). University of California at Los Angeles.

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs analytic, using two measurement models, the generalizability theory and many-facet Rasch measurement, within the context of performance assessment.* (Unpublished PhD dissertation). Pennsylvania State University, State College, PA.

Bachman, L. F. (1997). Generalizability theory. In C. Clapham & D. Corson (Eds.), *Encyclopedia of languages and education Volume 7: Language testing and assessment* (pp. 255 – 262). Dordrecht, Netherlands: Kluwer Academic.

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

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.

Banno, E. (2008). *Investigating an oral placement test for learners of Japanese as a second language.* (Unpublished PhD dissertation). Temple University, Philadelphia, PA.

Blok, H. (1999). Reading to young children in educational settings: A meta-analysis of recent research. *Language Learning, 49*(2), 343 – 371.

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

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

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

Brown, J. D. (1982). *Testing EFL reading comprehension in engineering English.* (Unpublished PhD dissertation). University of California at Los Angeles.

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

Brown, J. D. (1988). *1987 Manoa Writing Placement Examination: Technical Report #1*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1989). *1988 Manoa Writing Placement Examination: Technical Report #2*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1990a). *1989 Manoa Writing Placement Examination: Technical Report #5*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1990b). Short-cut estimates of criterion-referenced test consistency. *Language Testing, 7*(1), 77 – 97.

Brown, J. D. (1991). *1990 Manoa Writing Placement Examination: Technical Report #11*. Honolulu, HI: Manoa Writing Program, University of Hawai'i at Manoa.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas & C. Chapelle (Eds.), *A New Decade of Language Testing Research* (pp. 163 – 184). Washington, DC: TESOL.

Brown, J. D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*(2), 216 – 237.

Brown, J. D. (2005a). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D. (2005b). Statistics corner – Questions and answers about language testing statistics: Generalizability and decision studies. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(1), 12 – 16. Retrieved from http://jalt.org/test/bro_21.htm. [accessed Dec. 10, 2006].

Brown, J. D. (2007). Multiple views of L1 writing score reliability. *Second Language Studies* (Working Papers), *25*(2), 1-31.

Brown, J. D. (2008). Raters, functions, item types, and the dependability of L2 pragmatic tests. In E. Alcón Soler & A. Martínez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching and testing* (pp. 224 – 248). Clevedon, UK: Multilingual Matters.

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

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University.

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

Chiu, C. W.T. (2001). *Scoring performance assessments based on judgments: Generalizability theory*. Boston: Kluwer Academic.

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

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.

Gao, L., & Rodgers, T. (2007). Cognitive-psychometric modeling of the MELAB reading items. Paper presented at the National Council of Measurement in Education Conference, Chicago,IL.

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

Gerbil, A. (2010). Bringing reading-to-writing and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*, 100 – 117.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher, 5*, 3 – 8.

Goldschneider, J., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 51*, 1–50.

Jeon, E., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J. Norris & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching* (pp. 165 – 211). Philadelphia: John Benjamins.

Kim, Y.H. (2009). A G-theory analysis of rater effect in ELS speaking assessment. *Applied Linguistics, 30*(3), 435 – 440.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation of Japanese into English. *Language Testing, 21*(1), 1 – 27.

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.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied measurement in education, 2*(3), 195 – 205.

Lee, Y.-W. (2005) *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. TOEFL Monograph MS-28. Princeton, NJ: ETS.

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.

Lee, Y.-W, Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater.* TOEFL Research Report RR-81. Princeton, NJ: ETS.

Lee, Y.-W, & Kantor, R. (2005). *Dependability of ESL writing test scores: Evaluating prototype tasks and alternative rating schemes.* TOEFL Monograph MS-31. Princeton, NJ: ETS.

Lee, Y.-W, & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing, 7*(4), 353 – 385

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*, 158 – 180.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), Conversational interaction in second language acquisition: A series of empirical studies (pp. 407 – 452). Oxford: Oxford University.

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning, 53*, 123 – 163.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Molloy, H., & Shimura, M. (2005). An examination of situational sensitivity in medium-scale interlanguage pragmatics research. In T Newfields, Y. Ishida, M. Chapman, & M. Fujioka (Eds.), *Proceedings of the May 22 – 23, 2004 JALT Pan-SIG Conference* Tokyo: JALT Pan SIG Committee (pp. 16-32). Available online at www.jalt.org/pansig/2004/HTML/ShimMoll.htm. [accessed Dec. 10, 2006].

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417 – 528.

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3 – 50). Philadelphia: John Benjamins.

Norris, J. M., & Ortega, L. (2007). The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly, 41*(4), 805 – 815.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85 – 110.

Park, T. (2007). *Investigating the construct validity of the Community Language Program (CLP) English Writing Test.* (Unpublished PhD dissertation). Teachers College, Columbia University, New York, NY.

Rolstad, K., Mahoney, K., & Glass, G. (2005). Weighing the evidence: A meta-analysis of bilingual education in Arizona. *Bilingual Research Journal, 29*, 43 – 67.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1 – 20.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133 – 164). Philadelphia: John Benjamins.

Sahari, M. (1997). Elaboration as a text-processing strategy: A meta-analytic review. *RELC Journal, 28*(1), 15 – 27.

Sawaki, Y. (2003). A comparison of summarization and free recall as reading comprehension tasks in web-based assessment of Japanese as a foreign language. (Unpublished PhD dissertation). University of California at Los Angeles.

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.

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

Shavelson, R. J., & Webb, N. M. **(**1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-166.

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

Shin, S. (2002). Effects of subskills and text types on Korean EFL reading scores. *Second Language Studies* (Working Papers), *20*(2), 107-130. Retrieved from http://www.hawaii.edu/sls/uhwpesl/on-line_cat.html. [accessed Dec. 10, 2006].

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in testing of linguistic minorities. *Educational Measurement: Issues and Practice,* Spring, 13-22.

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.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239–261

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum.

Tang, X. (2006). Investigating the score reliability of the English as a Foreign Language Performance Test. (Unpublished PhD dissertation). Queen's University, Kingston, Ontario, Canada.

Taylor, A., Stevens, J., & Asher, W. (2006). The effects of explicit reading strategy training on L2 reading comprehension: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on second language learning and teaching (pp. 3-50).* Philadelphia: John Benjamins.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing,* 23(4), 411 – 440.

Van Weeren, J., & Theunissen, T. J. J. M. (1987). Testing pronunciation: An Application of generalizability theory. *Language Learning, 37*(1), 109 – 122.

Xi, X. (2003). *Investigating language performance on the graph description task in a semidirect oral test.* (Unpublished PhD dissertation). University of California at Los Angeles.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2) 251 – 286.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST)*. TOEFL iBT Research Report, TOEFLiBT-01. Princeton, NJ: ETS.

Yamamori, K. (2003). Evaluation of students' interest, willingness, and attitude toward English lessons: Multivariate generalizability theory. *The Japanese Journal of Educational Psychology, 51*(2), 195 – 204.

Yamanaka, H. (2005). Using generalizability theory in the evaluation of L2 writing. *JALT Journal, 27*(2), 169-185.

Yoshida, H. (2004). *An analytic instrument for assessing EFL pronunciation.* (Unpublished Ed.D. PhD dissertation). Philadelphia, PA: Temple University.

Yoshida, H. (2006). Using generalizability theory to evaluate reliability of a performance-based pronunciation measurement. (Unpublished ms). Osaka Jogakuin College.

Zhang, S. (2004). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. (Unpublished MA thesis). Ontario Institute for Studies in Education of the University of Toronto.

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing, 23*(3), 351 – 369.

Zhang, Y. (2003). Effects of persons, items, and subtests on UH ELIPT reading test scores. *Second Language Studies, 21*(2), 107-128. Retrieved from http://www.hawaii.edu/sls/uhwpesl/on-line_cat.html. [accessed Dec. 10, 2006].