A MULTIVARIATE GENERALIZABILITY ANALYSIS OF STUDENT STYLE QUESTIONNAIRE

By

YOUZHEN ZUO

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN EDUCATION

UNIVERSITY OF FLORIDA

2007

To my Mom

ACKNOWLEDGMENTS

First, I thank my advisor and committee chair Dr. David Miller for his invaluable guidance, feedback and encouragement throughout each semester. I also thank my committee member Dr. Thomas Oakland for providing me the data source and his quick feedback to my questions. Without their assistance, this document is impossible.

Next, I give my special thanks to Elaine Green and Linda Parsons who assist me in the track of my graduate study and their smiling faces make me feel like home in this department.

Last, I deeply thank my husband who always believes in me and supports me for my graduate career and my two children (Yuanze and Bill) who bring a lot of fun to my life.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts in Education

A MULTIVARIATE GENERALIZABILITY ANALYSIS OF STUDENT STYLE QUESTIONNAIRE

By

Youzhen Zuo

December 2007

Chair: David Miller
Major: Research and Evaluation Methodology

The Student Style Questionnaire (SSQ) is used to measure a student's temperament. In this study, the method of multivariate generalizability theory is applied to assess the reliability of the SSQ. In particular random effect variance and covariance components were estimated in the Generalizability (G) study. Meanwhile, generalizability coefficients for four subscales and the total scale also were estimated in the subsequent Decision (D) study. The results showed that the generalizability coefficients (reliability) were acceptable for the total scale and two of the subscales.

CHAPTER 1
INTRODUCTION

In education and psychology, tests are used to measure examinees' abilities and attitudes. One issue that test developers and users concern about is the reliability of the test. Reliability refers to the consistent replication of a measurement procedure across conditions, that is, a consistent score for an examinee would obtained over different test occasions, over parallel test forms, or over a set of raters. Reliability issue is so important that *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) required the test developers and users to obtain and report evidence concerning reliability and errors of measurement as a standard practice. Accordingly, investigators have developed methods that allow us to look into the measurement and target the error(s). To investigate the reliability, traditionally, we use Classical Test Theory (CTT), where an observed score is decomposed as a true score and an error score. However, one limitation of CTT is that only one undifferentiated error term can be specified at one time. Generalizability Theory (GT) was developed later to overcome this limitation and account for multiple sources of errors existed in the measurement simultaneously. In this thesis, CTT is briefly described as a comparison of method, however, our focus is on methodology using Generalizability Theory, especially, Multivariate Generalizability Theory (MGT), where not only multiple sources of errors can be specified but also reliability of a test (or profiles) including two or more subtests can be constructed. As a substantive illustration, we apply MGT to data from Student Style Questionnaire (SSQ) in later chapters.

Classical Test Theory

Classical test theory (CTT) is the most popular and simple method to assess the reliability of a measurement because it is easier to understand and conceptualize. Under CTT, an observed

score is conceptually decomposed as a true score and an error score; the agreement between the true score and the observed score is based on the reliability of the measurement. In common sense, the higher the reliability of a measure is, the closer the true score is to the observed score. Although we can never obtain a true score for an examinee, using so-called *parallel tests*, we can obtain the true score and observed score variances (Allen & Yen, 2002). As a result, the statistical representation of reliability is defined as the ratio of true score variance to the observed score variance. In addition, the reliability estimation in CTT addresses one source of error, such as test-retest, alternative forms, inter-rater, internal consistency, at a time (Crocker and Algina, 1986), which leads to different reliability coefficients in terms of different research designs. At the same time, comparing the different sources of errors existed in a measure and specifying the largest error is always difficult under the framework of CTT. In sum, with one and only one undifferentiated error term included in the model, CTT is limited to provide more insightful information on measurement error(s) that can help test developers and users adopt a better measurement design.

Generalizability Theory

To address the challenge coming from CTT, Cronbach, Gleser, Nanda and Rajaratnam (1972) first introduced Generalizbility Theory (GT) as a statistical theory for evaluating the reliability of measurements. Later, Shavelson and Webb further developed GT and made the theory more understandable with their published book—*Generalizability Theory: A Primer*. (1991). The theory reached its climax and was accepted by most researchers nowadays after R. L. Brennan produced his book—*Generalizability Theory* (2001a). Comparing with CTT, GT is more flexible and powerful. In particular, instead of decomposing an observed score as a true score and an error score, GT considers both systematic and unsystematic sources of error variations and disentangles them simultaneously, so the observed score can be decomposed into

as many possible effects as specified by the measurement design. For example, in a writing test where raters and prompts must be considered, an examinee's score can be decomposed into a grand mean in the population and universe, and seven other effects, due to person, rater, prompt, person-rater interaction, person-prompt interaction, rater-prompt interaction, and person-rater-prompt interaction. By examining all possible sources of error, a researcher can easily identify where large error sources come from and make appropriate decisions to decrease the error variance. As an extension of classical test theory, GT shares some concepts and assumptions with classical test theory. For instance, the *universe score* in GT has the same implication as the *true score* in classical test theory, errors are assumed to be uncorrelated and independent of true scores, samples selected and used to estimate the error variances are randomly selected from the population.

Many reliability studies utilizing Generalizability Theory (GT) is called Univariate Generalizability Theory (UGT), because only one *universe score* is associated with the object of measurement. For example, in math achievement test, each examinee only has one math score, that is, only one universe math score is associated with each person. Increasingly, however, datasets in the form of multiple subtests are more common to test developers and users. For example, in SAT test, each examinee has two scores representing verbal and math abilities, and a total score for the whole test. Of course, we could analyze the SAT data using UGT where the universe score of each examinee is regressed on his or her total score. However, we lose information about two specific subtests. In addition, if different set of items is used to measure each ability and the number of items is not equal, then, we have unbalance data problem, which leads to the difficulty of variance components estimation. Even worse, some measurement may not have a composite score, that is, they are only profiles for subscores. All these facilitated the

development of Multivariate Generalizability Theory (MGT), where two or more universe scores are associated with the object of measurement and covariance components in addition to variance components are taken into account. In sum, under certain circumstance, both UGT and MGT can analyze the same given data and the results from univariate and multivariate analysis can be similar, however, multivariate analysis provides more information that can be used by the test developers and users. The advantages of Multivariate GT over Univariate GT are obvious and will be described in detail in later chapter.

Student Style Questionnaire

In this study, we apply Multivariate Generalizability Theory (MGT) to evaluate the reliability of the Student Style Questionnaire (SSQ) Scale. The SSQ (Oakland, Glutting, Horton, 1996) includes 69 items designed to measure students' temperament on four bipolar subscales: *Extroverism/Introverism, Thinking/Feeling, Practical/Imaginative,* and *Organized/Flexible*. Students respond to 69 forced-choice questions related to real-life situations to express their individual style preferences. Each item is a brief description of a daily basis event, followed by two mutually exclusive alternative responses that indicate the student's preferred style. The Student Style Questionnaire measures preferences and students answer the questions in terms of what they like to do. The results can provide information for teachers, parents and counselors to assist students' leaning, vocational decisions and emotional counseling. The SSQ also can help in parent training, suggesting communication strategies for both parents and teachers.

CHAPTER 2
LITERATURE REVIEW

Generalizability Theory Overview

There is an extensive literature on the Generalizability Theory (GT) and its applications since 1972. Recent literature (Gao & Brennan, 2001; Yin, 2005; Solano-Flores, G. & Li, M. 2006; Clauser, B.E., Harik, P. & Margolis M. J. 2006) has referred most to two books: *Generalizability Theory: A Primer* (Shavelson & Webb, 1991) and *Generalizability Theory* (Brennan, 2001a). Both books introduced basic concepts and statistical models underlying GT. However, the book *Generalizability Theory* by Brennan (2001a) has the most comprehensive descriptions of GT.

At the beginning of the book, Brennan describes that GT is a combination CTT and ANOVA (analysis of variance), which helps us conceptualize the relationships among the three. The first chapter of the book introduces the framework of GT and gives definitions of the concept of GT. In the later chapters, Brennan demonstrates and compares GT analysis in the context of different research design, and addresses the methodological challenge such as unbalanced random effects designs and variability of statistics in GT. Topics are developed from simplicity to complexity along two main lines, which is Univariate Generalizability Theory (UGT) and Multivariate Generalizability Theory (MGT). In general, both UGT and MGT focus on fundamental concepts, different research designs and statistical computation of variance components. However, MGT has more centers of attention regarding to estimating covariance components and composite variance components. The book reduces complicated mathematical and statistical formulas, and uses a language that can be understood by the readers without much statistical background. Most important to researchers, the book describes the computer programs that are developed by Brennan (2001b) to simplify the calculation of GT.

The importance of GT lies in three reasons (Shavelson, 1991): First, GT allows the researchers to estimate the magnitude of multiple independent sources of error variance existed in a measure simultaneously. Second, the estimated variance components for a measure can be used to carry out *Decision Studies*, which means we can decide the degree of error variation according to our desired accuracy of measurement. Third, GT allows the estimation of test score reliability based on whether the scores will be used to make relative (norm-referenced test) or absolute (criterion-referenced test) decisions. In this sense, GT expands CTT in that reliability of scores depends on how we use these scores. In sum, GT provides a flexible and powerful tool to assess the reliability of a measure. It not only allows us to investigate but also to design reliable observations, utilizing the decomposed sources of variation in the measurement and then minimizing the measurement error(s) to reach an optimal design. However, all gains from GT are not obtained without "pain". The challenge in GT is to make a distinction among different measurement designs, which involve many conceptual and statistical issues. In this thesis, notations from Brennan's book *generalizability theory* are used to illustrate the conceptual and statistical issues.

Variance Components

Variance components refer to the variations existed in a measure; it is the most fundamental concept of GT. Usually, we notice that persons' scores on a measure are different; some higher, some lower. In common sense, at least, we can think of one reason why the scores differ – different people have different abilities or attitudes on the tested content, which is the variation due to persons. In fact, observed scores for examinees can be different due to other reasons besides person variation. For example, in an achievement test, an individual score on a particular item is affected by a person effect (p) (systematic differences among people's temperament, or *object of the measurement*), an item effect (i, variability due to items), and a residual including

the person-item interaction (pi) and other unspecified effect. As a result, an observed score for one individual on one item can be stated as:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu) \quad (2\text{-}1)$$

Where X_{pi} refers to the score for any person in the population on any item in the universe, μ refers to the grand mean score over all persons and items, μ_p refers to average score for person *p* over all items, μ_i refers to the average score for item *i* over all persons. The grand mean is constant for all people and has no variance. $(\mu_p - \mu)$ designates a person effect, $(\mu_i - \mu)$ an item effect, and the last the residual effect involving the interaction and all other sources of error not identified in this design. The last three terms are random effects.

Accordingly, the variance of those X_{pi} scores can be decomposed as,

$$\sigma^2(X_{pi}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(pi,e) \ . \quad (2\text{-}2)$$

where $\sigma^2(X_{pi})$ refers to the observed score variance, $\sigma^2(p)$ refers to the person variance, $\sigma^2(i)$ refers to the item variance and $\sigma^2(pi,e)$ refers to variance due to residuals. Thus, the variance of observed scores can be partitioned into independent sources of variation due to differences among persons, items, and the residual term, which includes the person-item interaction.

Universe of Admissible Observations and Facets

From the perspective of GT, a measurement is considered as a sample from a universe of admissible observations. All admissible observations have the same quality so that they can be interchanged for the purpose of making a decision about the measurement (Shavelson & Webb, 1991). The universe of admissible observations is made of facets, where each facet represents one source of variation in the measurement. The universe can have one or more facets, depending on how we conceptualize the measurement. In addition, each facet includes one or more similar conditions. For example, in a writing test, prompts are a possible facet. Each

prompt in the test can be considered a condition of this facet. An individual's score on a prompt is an observation. If each prompt is treated as a sample from a universe of prompts that are acceptable to test the writing ability, then the variation of the writing scores could come from the facet prompts besides differences in persons, which leads to the estimation of variance component for the facet prompts.

Generalizability (G) Studies and Decision (D) Studies

GT distinguishes G study and D study as two different but related procedures to investigate and design reliable observations. The purpose of a G study is to identify and decompose different sources of variations from the observed scores, while the purpose of a D study is to use the obtained variance components from G study to correct the error by collecting new sample from the prespecified universe of generalization and design the measurement for a particular purpose. Although both studies are involved with estimation of variance components, only D study allows us to manipulate the estimated variance components and change the measurement error(s). Therefore, G study is associated with the development of a measurement procedure, while D study is involved with application of the measurement procedure.

Generalizability Coefficient and Dependability Index

Generalizability (G) Coefficient and Dependability Index (phi coefficient) both represent the degree of agreements between universe score variance and observed score variance. However, G coefficient is different from Dependability Index in that: G coefficient is associated with a relative decision, where a decision of a score is based on relative standing of a person in the population (e.g., norm-referenced test); while Dependability Index is involved with a absolute decision, where a decision of a score is based on whether a person passes a standard or cutoff score (e.g., criterion-referenced test).

The Gneralizability coefficient in GT is analogous to the reliability coefficient in CTT. The Generalizability coefficient is defined as the ratio of universe score variance to the observed-score variance, which includes both the universe score variance and the relative error variance. For example, for an achievement test with items as the only facet, the relative error is

$$\sigma_{rel}^2 = \frac{\sigma_{pi,e}^2}{n_i} \tag{2-3}$$

where $\sigma_{pi,e}^2$ is the variance due to person by item interaction and unspecified error, n_i is the number of items in the measurement. The relative error (σ_{rel}^2) for an instrument is inversely proportionate to its number of items. Obviously, when the number of items increases, we can get small relative error; when the number of items decreases, we get large relative error.

The formula for calculating the generalizability coefficient is:

$$\mathrm{E}\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{rel}^2} \tag{2-4}$$

Where $E\rho^2$ refers to Generalizability coefficient, σ_p^2 refers to universe score variance, and σ_{rel}^2 refers to the relative error existed in the measurement. The generalizability coefficient shows how accurate a person's observed score can be generalized to his or her universe score (Shavelson & Webb, 1991). It reflects how much of the observed score variance that is due to universe score variance.

Similarly, Dependability index is defined as the ratio of universe score variance to the observed score variance, which includes universe score variance and absolute error variance.
In the same scenario of an achievement test mentioned above, the absolute error is

$$\sigma_{abs}^2 = \frac{\sigma_i^2 + \sigma_{pi,e}^2}{n_i} \tag{2-5}$$

where σ^2_{abs} refers to the absolute error, σ^2_i refers to the variance due to item, $\sigma^2_{pi,e}$ refers to the variance due to person by item interaction and unspecified error, and n_i refers to the number of items in the measurement. Notice that not only the residual variance (interaction and unidentified error) but also the items variance contributes to the absolute error. When the number of items increases, the absolute error decreases; and vise versa.

The formula for calculating the Dependability index is

$$\phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{abs}} \tag{2-6}$$

where ϕ refers to Dependability index, σ^2_p refers to the universe score variance, and σ^2_{abs} refers to the absolute error.

From equation 2-4 and 2-6, we notice that both G-coefficient and Dependability index involve the number of items in the measurement, therefore, we can determine how many items are necessary in a measurement in order to reach a specific ρ^2 or ϕ.

To correctly specify the relative and absolute error in the measurement is important for calculating the right G coefficient and dependability index. Usually, relative error only involves error variances that only interact with the person (object of measurement). In contrast, absolute error involves all variances except for the universe score variance. This difference is demonstrated with the Venn diagram in Figure 2-1.

Random and Fixed Facets

Whether a facet is random or fixed depends on how we define the conditions of the facet. Suppose the conditions of a facet in a measurement are randomly sampled from conditions in the universe and will be generalized to the other conditions not included in the sample. In this case, the conditions in the measurement represent a sample from all possible conditions. Therefore, the

facet is a *random* one. On the contrary, a facet is a *fixed* one when the conditions of the facet in a measurement are the only conditions in the universe and could not be generalized to other conditions. The importance of recognizing a fixed or random facet is in that, the variance associated with a fixed facet is not regarded as an error variance component. Therefore, when we calculate the reliability coefficients for the measurement, variance due to a fixed facet should not be included in relative or absolute error but as a part of the universe score variance. However, the fixed facet presents a problem when GT treats it by averaging over the conditions of the facet, which does not always make a conceptual sense. The solution that Shavelson & Webb (1991) suggested is to conduct separate G studies within each condition of the fixed facet. This suggestion is exactly aligned with MGT that we will present in later section.

Crossed and Nested Facets

Crossed facet happens when every condition of one facet interacts with every condition of other facet. For example, each individual responds to all the items in the achievement test. The design is a *crossed* one and can be denoted by p×i, where p represents person and i represents items. In contrast, nested facet happens when different set of conditions of a facet interacts with one and only one condition of another facet. For example, in SAT test, different sets of items are associated with either verbal subtest or math subtest but not both; therefore, the items are nested within the subtest facet. We can denote this nested design as i:t, where t represents the facet subtest. In addition, a design of a measure may include both crossed and nested facets at the same time. For example, each individual responds to all the items, and two or more items appear with one and only one condition (e.g., subtest) of an achievement test. The notational form of this design may be expressed as p×(i:t). The importance of recognizing the difference between

crossed and nested facet is in that, estimation of variance components from a *nested* design is different from a *crossed* design, which directly leads to different reliability estimation.

Multivariate Generalizability Theory Overview

For the last few decades, test developers and users have attempted to investigate the reliability of a measurement where responses of multiple subtests (or profiles) are obtained for each object of measurement. Such data have the following characteristics: (1) each examinee (object of measurement) has two or more universe scores representing subtests or profiles; (2) The conditions of subtests (or profiles) are fixed, that is, the selected conditions are our interest and will not be generalized to other conditions. (3) The number of items in each condition of the subtests (or profiles) is not the same, which means the data are unbalanced. Furthermore, the researchers concern about not only each universe score but also the composite (or profiles) of universe scores for the whole test. Multivariate generalizability theory (MGT), in contrast with univariate generalizability theory (UGT), was developed to meet the challenge (Rajaratnam, 1965; Shavelson & Webb, 1991; Brennan, 2001a).

MGT is not complete without comparing with UGT. The difference between MGT and UGT can be described like this: MGT involves with two or more universe scores for the object of measurement at the same time, while UGT involves with only one universe score for the object of measurement at a time. In this sense, multivariate analysis of a specific dataset can be constructed based on multiple univariate analyses in a row. More importantly, multivariate analysis account for not only variance components like univariate analysis does, but also covariance components between the universe scores that univariate analysis cannot do. This powerful function of multivariate analysis allows us to investigate and design reliable observations both at each universe score level and composite score or profiles level.

However, each multivariate design can have a counterpart in a univariate design, which means, logically, that any data can be performed with univariate GT. The choice of multivariate GT over univariate GT depends on the complexity of the data and what kind of information that we want to derive. Brennan (2001) recommended performing a full multivariate analysis if there is a fixed facet in the research design. In his book *Generalizability Theory*, Brennan discussed the problems of analyzing unbalanced data, where the sample sizes in each condition of a facet are not equal. Unbalanced data creates complexity when we want to decompose the variance components. One way to reduce the complexity of unbalanced data in univariate analysis is to analyze the data under the framework of multivariate design if possible. Take the SSQ data as an example: it is reasonable that different sample sizes of items are distributed to four temperament scales in SSQ test with more items in one scale and fewer items in the other scale. Multivariate design avoids the problem of unbalanced data by analyzing four parallel univariate designs. In the end, each univariate design has balanced data under four levels of fixed facet temperament.

In addition, Haertel (2006) pointed out two disadvantages of using univariate analysis for data containing fixed facet. First, variance components are forced to be the same for observed scores for the levels of the fixed facet. Second, universe score represents an equally weighted composite of scores on the levels of the fixed facet, which is not always true. Consequently, some information about the scores cannot be derived from univariate analysis.

In sum, the advantages of multivariate GT over univariate GT are: a) multivariate GT reduces the complexities and ambiguities in terms of unequal numbers of items within fixed facets if univariate analysis is used; b) estimations of variance and covariance components can be alienated in a multivariate analysis, but not in univariate analysis (Brennan, 2001a); (c) estimate observable correlations, or universe-score and error correlations for various D study

designs (Brennan, 2001a); (d) estimate the reliability of profiles of scores using multiple regression of universe scores on the observed scores in the profile (Brennan, 2001a, Cronbach et al., 1972); or (e) produce a composite of scores with maximum generalizability (Shavelson & Webb, 1981). For all of these purposes, multivariate GT can be a better choice.

A later chapter will present the methods of how to compute variance and covariance components, disattenuated correlation, and Generalizability coefficient and Dependability index under the framework of multivariate GT using the actual SSQ data as an illustration.

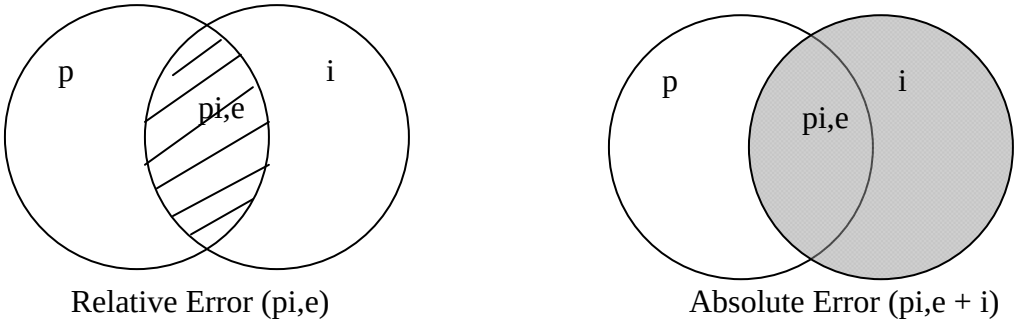

Figure 2-1. Relative and absolute error for a random p × i design

CHAPTER 3
RESEARCH DESIGN AND METHODOLOGY

Student Styles Questionnaire Data

The Student Styles Questionnaire (SSQ; Oakland, Glutting, & Horton, 1996) is an instrument designed to measure children's temperament. Temperament refers to people's consistent attitudes, preferences, affect, and styles of behaving. In education, a student's temperament influences the ways that students learn and associate in school. It is important for teachers and parents to understand students' temperament so that effective and appropriate teaching methods can be applied to the students. There are four pairs of temperament styles: *extroversion-introversion* (EI), *practical-imaginative* (PM), *thinking-feeling* (TF) and *organized-flexible* (OL) (Oakland et al., 1996). Each of the four bipolar styles contrasts aspects of a student's preferences.

SSQ includes 69 self-report temperament items; each item has two options in a forced-choice response format. It was designed to be used with children ages 8 through 17. Each temperament style was measured with different set of items—23 items measure EI style, 16 items PM style, 10 items TF style, and 26 items OL style. One feature of coding the responses is that, if the response of the student belongs to extrovert, thinking, organized or practical style, the score is positive; otherwise, the score is negative. One example of the items is "After school I like to do something (alone/with friends)". If the response of the student is "with friends" (extrovert-type), the score for the item is 3; if "along" (introvert-type), the score is –2. (Note that 6 out of 69 items measure two different styles simultaneously; for example, item 41 measures OL as well as PM. We ignored the complexity of 6 items crossed with two levels of temperament for convenience of illustration).

The SSQ data used in this study were acquired through a random sample representative of the 1990 U.S. Bureau of the Census data. Total 7,902 children from the public and private schools with the age range of 8 through 17 completed the questionnaires (6.7% age 8, 8.2% age 9, 11.8% age 10, 14.9% age 11, 14.2% age 12, 14.7% age 13, 11.7% age 14, 8.1% age 15, 5.2% age 16, and 4.6% age 17). There were 3952 (50%) female and 3950 (50%) male in the sample. The data were collected from four different regions: 1976 (25% of total) students were from the North Central, 1692 (21.4% of total) students from the West, 2785 (35.2% of total) from the South, and 1449 (18.3%) from the Northeast. There were 5547 (70.2%) white students, 868 (11%) Hispanic students, 1194 (15.1%) black students, and 293 (3.7%) other ethnic students.

Multivariate Design and Research Questions

SSQ has four temperament scales: *extroversion-introversion* (EI), *practical-imaginative* (PM), *thinking-feeling* (TF) and *organized-flexible* (OL). Different set of items is associated with one and only one content area, which means, the items are nested within the fixed facet. Temperament is fixed facet here because we have no intention to generalize four types of temperament to other conditions of content. When the students take the SSQ, four subtest scores are reported to them. Each subtest score (x_{pi}) for an examinee can be decomposed into effects due to persons (p), items (i), person-by-item interaction (pi) or residual effects, and a grand mean (μ) for the content. This is analogous to a univariate analysis associated with each level of the fixed content facet. Finally, we have four such univariate designs in a row, which form the multivariate design for SSQ.

The notational form of multivariate design of SSQ is $p^{\bullet} \times i^{\circ}$, where p stands for students and i stands for items. The solid circle means that students are crossed with multivariate fixed facet, while empty circle means that items are nested within each condition of the multivariate fixed

facet. In addition, the existence of covariance component depends on whether there is a random facet linked with the fixed facet. “Linked” has the same meaning with “crossed”. As a result, the expected values of error covariance components for students (p) are not zero, while the expected values of error covariance components for items (i) are zero.

Applying MGT for assessment of the SSQ data is important in that different aspects or facets of individual score differences can be quantified and used in future decision-making. In particular, the following questions will be examined in our study:

1. What is the magnitude of error variances in SSQ?
2. Are the four temperaments in SSQ correlated?
3. What are the generalizability and dependability coefficients for each subscale in SSQ?
4. What are the generalizability and dependability coefficients for SSQ profiles?

Multivariate Generalizability Studies

Multivariate analysis of SSQ data consists of two steps: multivariate generalizability (G) studies and multivariate decision (D) studies. In the first step, multivariate G studies focus on estimating variance and covariance components, and disattenuated correlations between temperament styles. In the second step, multivariate D studies (in next section) focus on minimizing measurement error(s) given the information obtained from the first step, and calculating appropriate G or phi coefficients.

Estimating Variance Components

The multivariate generalizability study $p^{\bullet} \times i^{\circ}$ design is used to estimate the variance and covariance components for the SSQ data. Letting v_1 to v_4 stand for the four temperament styles (EI, TF, PM, and OL) for the SSQ, p stand for persons or students, and i stand for items or questions, a linear model can be used to describe the observed scores for each of the temperament scales. Taking v_1 (EI) as an example:

$$X_{piv1} = \mu_{v1} + \nu_{pv1} + \nu_{iv1} + \nu_{piv1}, \quad (3\text{-}1)$$

in which the ν terms represent effects for ν_1, and μ_{v1} is the grand mean for the universe score for EI. An examinee's observed score X_{piv1} for EI can be decomposed into effects due to persons (ν_{pv1}), items (ν_{iv1}), person-by-item interaction or residual effects (ν_{piv1}), and a grand mean for the level of temperament (μ_{v1}).

Similar equations can be used to describe observed scores for TF, PM and OL. That is, there is a univariate p × i design associated with each level of temperament ν. The superscript filled circle for $p(^{\bullet})$ indicates that persons have scores on all four temperament scales. The superscript empty circle for i ($^{\circ}$) indicates that items are nested within each level of temperament; that is, each set of items belongs to only one level of temperament. The fixed levels of ν are linked in the sense that the same group of students responds to all items for all levels. The Venn diagram for multivariate $p^{\bullet}\times i^{\circ}$ design is showed in Figure 3-1, where v is the fixed facet temperament.

Similarly, the variance of the scores for one scale (i.e., EI) over the population of students and items in the universe of admissible observations can be expressed as below:

$$\sigma^2(X_{pi}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(pi) \quad (3\text{-}2)$$

Variance components cannot be observed directly. To obtain the variance component, we need ANOVA procedure to obtain sum squares and expected mean squares (EMS) first. After obtaining the EMS from ANOVA procedure, we obtain the variance components estimators using following equations:

$$\hat{\sigma}^2(p) = \frac{MS(p) - MS(pi)}{n_i} \quad (3\text{-}3)$$

$$\hat{\sigma}^2(i) = \frac{MS(i) - MS(pi)}{n_p} \quad (3\text{-}4)$$

$$\hat{\sigma}^2(pi) = MS\ (pi) \quad (3\text{-}5)$$

where n_i is the number of items, n_p is the number of students.

Estimating Covariance Components

An unbiased estimator of the covariance between universe scores for the linked facets is calculated according to the following equation:

$$\hat{\sigma}_{vv'}(p) = S_{vv'}(p) = \frac{n_p}{n_p - 1}\left(\frac{\sum_p \overline{X_{pv}}\,\overline{X_{pv'}}}{n_p} - \overline{X_v}\,\overline{X_{v'}}\right) \quad (3\text{-}6)$$

where $S_{vv'}(p)$ is the observed covariance between two universe scores, $\overline{X_{pv}}$ refers to mean score for specific person in one specific type of temperament average across items, $\overline{X_v}$ refers to mean score for one specific type of temperament average across all persons and items, and n_p is the number of students.

Variance and Covariance Matrix

Variance and covariance components for the SSQ $p^{\bullet} \times i^{\circ}$ design are presented here using matrix conventions:

$$\Sigma_p = \begin{pmatrix} \sigma_1^2(p) & \sigma_{12}(p) & \sigma_{13}(p) & \sigma_{14}(p) \\ \sigma_{21}(p) & \sigma_2^2(p) & \sigma_{23}(p) & \sigma_{24}(p) \\ \sigma_{31}(p) & \sigma_{32}(p) & \sigma_3^2(p) & \sigma_{34}(p) \\ \sigma_{41}(p) & \sigma_{42}(p) & \sigma_{43}(p) & \sigma_4^2(p) \end{pmatrix} \quad (3\text{-}7)$$

$$\Sigma_i = \begin{pmatrix} \sigma_1^2(i) & & & \\ & \sigma_2^2(i) & & \\ & & \sigma_3^2(i) & \\ & & & \sigma_4^2(i) \end{pmatrix} \quad (3\text{-}8)$$

$$\Sigma_{pi} = \begin{pmatrix} \sigma_1^2(pi) & & & \\ & \sigma_2^2(pi) & & \\ & & \sigma_3^2(pi) & \\ & & & \sigma_4^2(pi) \end{pmatrix} \quad (3\text{-}9)$$

In these matrices, each column represents one level of the fixed facet of SSQ in the following order: EI, PM, TF, and OL. Elements on the diagonal are estimated variance

components, and elements on the off-diagonal are estimated covariance components. Because persons are linked, covariance components can be estimated between pairs of four levels for the object of measurement. Meanwhile, items and person-by-item interaction terms are not linked. Thus covariance components between pairs of four levels are zero and therefore are not listed.

Disattenuated Correlations

In measurement, disattenuated correlations refer to correlation between two measures accounting for measurement error, which is in contrast with weakened correlations because of measurement error. Disattenuated correlations for pairs of universe scores of four levels of SSQ through the object of measurement (students) can be estimated. For example, the disattenuated correlation coefficient between two temperament scales (v and $v^{'}$) for the universe score (person) is

$$\rho_{vv'}(p) = \frac{\sigma_{vv`}(p)}{\sqrt{\sigma_{v}^{2}(p)\sigma_{v`}^{2}(p)}} \qquad (3\text{-}10)$$

where $\sigma_{vv`}(p)$ is the covariance of two universe scores, $\sigma_{v}^{2}(p)$ is one of the universe score variances.

Multivariate Decision Study

Estimates of variance and covariance components for the decision (D) study are pretty straight forward on the basis of the G study results. Σ_p is unchanged, and Σ_I and Σ_{pI} can be obtained by dividing the diagonal elements in Σ_i and Σ_{pi} by the number of decision study items ($n_i^{'}$) within each level.

The three D study variance-covariance matrices are as below,

$$\Sigma_p = \begin{pmatrix} \sigma_1^2(p) & \sigma_{12}(p) & \sigma_{13}(p) & \sigma_{14}(p) \\ \sigma_{21}(p) & \sigma_2^2(p) & \sigma_{23}(p) & \sigma_{24}(p) \\ \sigma_{31}(p) & \sigma_{32}(p) & \sigma_3^2(p) & \sigma_{34}(p) \\ \sigma_{41}(p) & \sigma_{42}(p) & \sigma_{43}(p) & \sigma_4^2(p) \end{pmatrix} \quad (3\text{-}11)$$

$$\Sigma_I = \begin{pmatrix} \sigma_1^2(i)/n_{i1} & & & \\ & \sigma_2^2(i)/n_{i2} & & \\ & & \sigma_3^2(i)/n_{i3} & \\ & & & \sigma_4^2(i)/n_{i4} \end{pmatrix} \quad (3\text{-}12)$$

$$\Sigma_{pI} = \begin{pmatrix} \sigma_1^2(pi)/ n_{i1} & & & \\ & \sigma_2^2(pi)/ n_{i2} & & \\ & & \sigma_3^2(pi)/ n_{i3} & \\ & & & \sigma_4^2(pi)/ n_{i4} \end{pmatrix} \quad (3\text{-}13)$$

Relative and absolute error variance-covariance matrices can then be obtained:

$$\Sigma_\delta = \Sigma_{pI} \quad (3\text{-}14)$$

$$\Sigma_\Delta = \Sigma_I + \Sigma_{pI} \quad (3\text{-}15)$$

Equations 3-14 and 3-15 are diagonal matrices for the $p^\bullet \times I^\circ$ design. The square roots of relative and absolute error variance are the relative and absolute standard error of measurements (SEMs). Also, from Equation 3-12 to 3-15, we can see that the relative and absolute error variances are directly influenced by the numbers of items in each level of the fixed facet. In particular, when the number of items increases, the relative and absolute errors decrease. As a result, the G coefficient and dependability index increase.

SSQ Profiles

SSQ profiles include four temperament scales. A composite score as sum of subscores of four temperament scales for a student does not make sense here, however, we still can assess the SSQ profiles as one unitary instrument in terms of composite universe score variance and composite relative and absolute errors. Brennan (2001a) recommends three weights to sum the variance and covariance components, and one of them is priori weights. The rationale of a priori weight is that: an appropriate composite universe score variance is a weighted average of the

universe score variance for each level of a fixed facet, where the weights are proportional to the number of items in each level. Such a weight vector can be defined as

$$W` = [\frac{n_{i1}^{'}}{n_{+}^{'}} \quad \frac{n_{i2}^{'}}{n_{+}^{'}} \quad \frac{n_{i3}^{'}}{n_{+}^{'}} \quad \frac{n_{i4}^{'}}{n_{+}^{'}}], \tag{3-16}$$

where $n_{i1}^{'}$ to $n_{i4}^{'}$ are the number of items for each level of the fixed facet, and $n_{+}^{'}$ is the total number of items.

Composite Universe Score Variance, Relative and Absolute Score Variance

Using the weight vector w`, the variance for the composite universe score is

$$\sigma^2_C(p) = \sum_v w_v^2 \sigma_v^2(p) + \sum_{v \neq v'} \sum w_v w_{v'} \sigma_{vv'}(p) \tag{3-17}$$

where C is the composite score, *w* is the weight. Similarly, estimated relative and absolute error variances for the composite score are

$$\sigma^2_C(\delta) = \sum_v w_v^2 \sigma_v^2(\delta) = \sum_v \frac{w_v^2}{n_{iv}} \sigma_v^2(pi) \tag{3-18}$$

$$\sigma^2_C(\Delta) = \sum_v w_v^2 \sigma_v^2(\Delta) = \sum_v \frac{w_v^2}{n_{iv}} \left[\sigma_v^2(i) + \sigma_v^2(pi)\right] \tag{3-19}$$

Square roots of relative and absolute error variances for the composite are the relative and absolute standard error of measurement for the SSQ profiles.

G-Coefficient and Dependability Index for SSQ Profiles

A multivariate generalizability coefficient ($E\rho^2_C$) for the profiles can be defined as the ratio of composite universe score variance to itself plus composite relative error variance:

$$E\rho^2_C = \frac{\sigma_C^2(p)}{\sigma_C^2(p) + \sigma_C^2(\delta)} \tag{3-20}$$

A generalizability coefficient is often used for making norm-referenced interpretations. Similarly, a multivariate dependability index (phi coefficient) can be defined as the ratio of composite universe score variance to itself plus composite absolute error variance.

$$\phi_C = \frac{\sigma_C^2(p)}{\sigma_C^2(p) + \sigma_C^2(\Delta)} \tag{3-21}$$

A dependability index is often used for making criterion-referenced interpretation.

Data Analysis Software for This Study

All computations were performed using the SPSS 14.0 package and EXCEL. The sums of squares and expected mean squares (EMS) were computed using the repeated measures option of the GLM procedure in SPSS. EXCEL was used to calculate the estimated error variances from EMS, covariance components and disattenuated correlations between universe scores and all coefficients.

The availability of computer programs allow for a more convenient way to do multivariate data analysis by mGENOVA (Brennan 2001b). There is a detailed description of mGENOVA in Brennan' book *Generalizability Theory* (2001a). One advantage of mGENOVA program is to produce output for large size data set in a few minutes. However, this study will not use mGENOVA program.

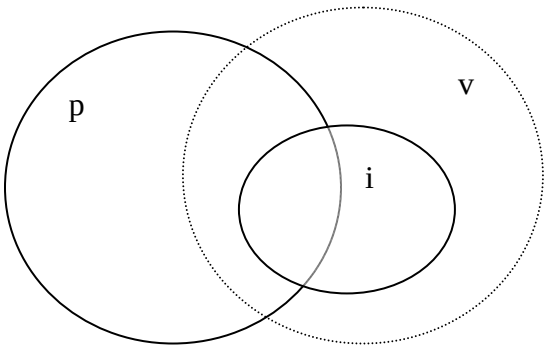

Figure 3-1. Representation for a multivariate $p^{\bullet} \times i^{\circ}$ design.

CHAPTER 4
RESULTS

In this study, the dataset was divided into eight random cases, with sample size for first seven cases is 1000, and the last case 902. The reasons for doing this are a) we want to examine whether eight cases representing eight random samples are consistent with each other in terms of magnitude of error variances and b) with large dataset (i.e. total 7902 students in the SSQ data), SPSS usually is "out of memory" to produce the results.

Multivariate Generalizability Study Results

Items are the only random facet in this multivariate generalizability analysis of SSQ data. Students are the objects of measurement. The items were presented to all the students as a crossed design. In addition, students were crossed with items that were nested within each level of the fixed facet temperament. Therefore, the complete design in this multivariate study is denoted as: $p^{\bullet} \times i^{\circ}$, where p refers to persons effect and i the items. The covariance components design is for p and variance components design is for i and $p \times i$. Since the fixed facet of this multivariate design has four levels, the multivariate design involves 4 variance components designs and 4(4-1)/2 =6 covariance components designs.

Table 4-1 through Table 4-4 provides mean squares and estimated variance components for EI, PM, TF and OL in case 1 SSQ data using equation 3-3 to 3-5, and Table 5-5 provides covariance components for pairs of scales in Case 1 SSQ data using equation 3-10.

Estimated variance and covariance components for eight samples of SSQ following the procedure of Case 1 are listed in Table 4-6. Each sample was administered to 1000 students (with the exception of the last case) in a random-groups design. Also provided are the means and standard deviations of the estimated variance and covariance components. According to Brennan (2001a), the standard deviations are empirical estimates of the standard errors, without making

normality assumptions. The relatively small magnitude of this standard error is reflected by the similarity of the eight estimates.

As shown in Table 4-6 and 4-7, variance and covariance component averages for all eight samples were similar in magnitude. Estimated variance components for p were relatively small (0.28 to 0.80). Estimated variance components for i were similar to those for p. Estimated variance components for pi were the largest, (eight times larger than variance component estimates for p and i), which indicates that relative ranking of different students tended to vary across different items.

Among the four scales of temperament, estimated variance components for persons were the largest for TF and least for PM. On average, TF has the largest variance components for persons (.69), PM has the smallest variance components for persons (.31). Estimated covariance components for p were small compared to the estimated variance components for p. The components of covariance for persons reflect the underlying correlation among the temperament components. On average, PM-TF has the largest covariance component (.15), EI-PM has the smallest covariance component (-.01), which indicates that persons who have p*ractical* temperament style are also more likely to have *thinking* temperament style, and persons who have *imaginative* style more likely to have *feeling* temperament style. However, for persons are *extrovert* or *introvert* temperament style, there is no difference in terms of preferences to *practical* or *imaginative* temperament style. The estimated variance components for items for OL were the largest and for TF were the smallest, which indicates that item difficulty for OL varied the most. Estimated variance components for pi for TF were the largest.

Using the means of the estimated variances and covariances in Table 4-6 and 4-7, the G study matrices are

$$\hat{\sum} p = \begin{pmatrix} .52 & -.01 & .12 & -.05 \\ -.01 & .31 & .15 & .13 \\ .12 & .15 & .69 & -.09 \\ -.05 & .13 & -.09 & .43 \end{pmatrix} \quad (4\text{-}1)$$

$$\hat{\sum} i = \begin{pmatrix} .59 & & & \\ & .46 & & \\ & & .44 & \\ & & & .68 \end{pmatrix} \quad (4\text{-}2)$$

$$\hat{\sum} pi = \begin{pmatrix} 4.73 & & & \\ & 3.68 & & \\ & & 4.89 & \\ & & & 4.16 \end{pmatrix} \quad (4\text{-}3)$$

Multivariate Decision Study Results

Using D study sample sizes of $n_{i1} = 23$, $n_{i2} = 16$, $n_{i3} = 10$, and $n_{i4} = 26$, (same as the G study sample size) the D study estimated variance-covariance matrices are

$$\hat{\sum} p = \begin{pmatrix} .52 & -.01 & .12 & -.05 \\ -.01 & .31 & .15 & .13 \\ .12 & .15 & .69 & -.09 \\ -.05 & .13 & -.09 & .43 \end{pmatrix} \quad (4\text{-}4)$$

$$\hat{\sum} I = \begin{pmatrix} .026 & & & \\ & .029 & & \\ & & .044 & \\ & & & .026 \end{pmatrix} \quad (4\text{-}5)$$

$$\hat{\sum} pI = \begin{pmatrix} .21 & & & \\ & .23 & & \\ & & .49 & \\ & & & .16 \end{pmatrix} \quad (4\text{-}6)$$

It follows that the estimated universe score, relative error, and absolute error matrices are

$$\hat{\sum} p = \begin{pmatrix} .52 & \textit{-.02} & \textit{.20} & \textit{-.11} \\ -.01 & .31 & \textit{.32} & \textit{.36} \\ .12 & .15 & .69 & \textit{-.17} \\ -.05 & .13 & -.09 & .43 \end{pmatrix} \quad (4\text{-}7)$$

$$\hat{\sum} \delta = \begin{pmatrix} .21 & & & \\ & .23 & & \\ & & .49 & \\ & & & .16 \end{pmatrix} \quad (4\text{-}8)$$

$$\hat{\sum} \Delta = \begin{pmatrix} .236 & & & \\ & .259 & & \\ & & .534 & \\ & & & .186 \end{pmatrix} \quad (4\text{-}9)$$

where the italicized values in the upper diagonal positions of $\hat{\sum} p$ are disattenuated correlations for universe scores, the values in the diagonal positions are variance components. The first column (or row) of the matrix represents EI, second PM, third TF and fourth OL. The range of the correlation is from negative .17 to positive .36, which indicates some types of temperament are correlated with each other to some degree.

The G coefficient and the dependability index Φ coefficient within each level of the fixed facet temperament can be obtained using equation 2-4 and 2-6; the results are in the following matrix in the order of EI, PM, TF and OL.

$$\begin{pmatrix} \mathrm{E}\hat{\rho}^2 \\ \hat{\Phi} \end{pmatrix} = \begin{pmatrix} .71 & .57 & .58 & .73 \\ .69 & .54 & .56 & .70 \end{pmatrix} \quad (4\text{-}10)$$

To have more choices of Generalizability coefficient and dependability index for each subscale of temperament, we carry out several D-studies using different items other than the items used in SSQ subscales. Table 4-7 shows the generalizability coefficients (G-coefficients), as well as the dependability coefficients, associated with the various items, beginning with the original G-study sample sizes for EI, PM, TF and OL.

Multivariate G and D Study Results for SSQ Profiles

To obtain universe scores variance and error variance of SSQ profiles, we assume that the numbers of items that contribute to each SSQ scale are reflective of the relative importance for the universe of generalization intended by SSQ. Under this assumption, a prior weights are w_1 = 23/75 = .31, w_2 =16/75 = .21, w_3 = 10/75 = .13, and w_4 = 26/75 = .35. Using these weights, estimated composite variances for the universe scores using equation 3-17 is shown below

$$\hat{\sigma}_C^2(p) = (.31)^2*(.52) + (.21)^2*(.31) + (.13)^2*(.69) + (.35)^2*(.43) + 2*(.31)(.21)(-.01) + 2*(.31)(.13)(.12) + 2*(.31)(.35)(-.05) + 2*(.21)(.13)(.15) + 2*(.21)(.35)(.13) + 2*(.13)(.35)(-.09) = .14$$

The estimated composite relative error variance using equation 3-18 is shown below

$$\hat{\sigma}_C^2(\delta) = (.31)^2*(.21) + (.21)^2*(.23) + (.13)^2*(.49) + (.35)^2*(.16) = .06,$$

The estimated composite absolute error variance using equation 3-19 is shown below

$$\hat{\sigma}_C^2(\Delta) = (.31)^2*(.236) + (.21)^2*(.259) + (.13)^2*(.534) + (.35)^2*(.186) = .07,$$

The Generalizability coefficient for the composite score of overall temperament in this multivariate G study can be calculated using equation 3-20. Since the true variances are always unknown, those estimated variance components are plugged in the equation. The calculations are shown below

$$E\hat{\rho}_C^2 = \frac{.14}{.14+.06} = .70$$

This indicates that about 70% of the variability in individuals' scores was systematic and attributable to the universe score. Such interpretation is similar to that for the reliability coefficient in CTT since CTT concerns the relative standing of individuals.

The dependability index was calculated using equation 3-21:

$$\hat{\Phi}_C = \frac{.14}{.14+.07} = .67$$

This coefficient should be used when absolute decisions are to be made using this measurement.

Table 4-1. Anova for G study p×i design for EI in case 1

Source of Variation	Df	Mean Squares	Estimated Variance Component
p	999	16.76	.52
i	22	567.77	.56
p×i	21978	4.76	4.76

Note: EI = extrovert-introvert, $n_p = 1000$, $n_i = 23$.

Table 4-2. Anova for G study p×i design for PM in case 1

Source of Variation	Df	Mean Squares	Estimated Variance Component
p	999	8.68	.31
i	15	401.32	.40
p×i	14985	3.72	3.72

Note: PM = practical-imaginative, $n_p = 1000$, $n_i = 16$.

Table 4-3. Anova for G study p×i design for TF in case 1

Source of Variation	Df	Mean Squares	Estimated Variance Component
p	999	12.08	.72
i	9	419.11	.41
p×i	8991	4.87	4.87

Note: TF = thinking-feeling, $n_p = 1000$, $n_i = 10$.

Table 4-4. Anova for G study p×i design for OL in case 1

Source of Variation	Df	Mean Squares	Estimated Variance Component
p	999	16.03	.45
i	25	604.07	.60
p×i	24975	4.21	4.21

Note: OL = organized-flexible, n_p = 1000, n_i = 26.

Table 4-5. Variance and covariance components for universe Scores of SSQ data in case 1

	EI	PM	TF	OL
EI	.52	-.01	.08	-.07
PM	-.01	.31	.16	.09
TF	.08	.16	.72	-.09
OL	-.07	.09	-.09	.45

Table 4-6. The G study variance and covariance components for person in SSQ

	Person									
Case	$\hat{\sigma}_1^2(p)$	$\hat{\sigma}_2^2(p)$	$\hat{\sigma}_3^2(p)$	$\hat{\sigma}_4^2(p)$	$\hat{\sigma}_{12}(p)$	$\hat{\sigma}_{13}(p)$	$\hat{\sigma}_{14}(p)$	$\hat{\sigma}_{23}(p)$	$\hat{\sigma}_{24}(p)$	$\hat{\sigma}_{34}(p)$
1	.52	.31	.72	.45	-.01	.08	-.07	.16	.09	-.09
2	.49	.32	.64	.43	-.04	.09	-.06	.15	.13	-.09
3	.46	.29	.65	.41	-.03	.12	-.02	.16	.15	-.10
4	.62	.33	.67	.42	-.01	.12	-.05	.15	.11	-.13
5	.46	.31	.78	.44	-.03	.15	-.08	.09	.19	-.05
6	.65	.30	.56	.42	-.03	.15	-.07	.12	.11	-.12
7	.56	.36	.80	.41	.004	.16	-.04	.19	.13	-.10
8	.43	.28	.67	.44	.03	.10	-.02	.16	.13	-.03
Mean	.52	.31	.69	.43	-.01	.12	-.05	.15	.13	-.09
SE	.08	.025	.077	.012	.023	.026	.022	.03	.026	.034

Note: In variance and covariance notation (i.e., $\hat{\sigma}_1^2(p)$), 1 = EI, 2 = PM, 3 = TF, 4 = OL.

Table 4-7. The G study variance components for item and person-by-item in SSQ

	Item				Person × Item			
Case	$\hat{\sigma}_1^2(i)$	$\hat{\sigma}_2^2(i)$	$\hat{\sigma}_3^2(i)$	$\hat{\sigma}_4^2(i)$	$\hat{\sigma}_1^2(pi)$	$\hat{\sigma}_2^2(pi)$	$\hat{\sigma}_3^2(pi)$	$\hat{\sigma}_4^2(pi)$
1	.56	.40	.41	.60	4.76	3.72	4.87	4.21
2	.63	.47	.50	.66	4.74	3.68	4.91	4.16
3	.57	.44	.43	.70	4.82	3.74	4.89	4.11
4	.64	.45	.51	.77	4.62	3.69	4.85	4.14
5	.58	.54	.44	.68	4.81	3.63	4.73	4.14
6	.58	.47	.43	.65	4.59	3.67	5.09	4.28
7	.65	.51	.46	.73	4.66	3.59	4.79	4.14
8	.52	.39	.31	.61	4.84	3.75	4.96	4.11
Mean	.59	.46	.44	.68	4.73	3.68	4.89	4.16
SE	.043	.046	.061	.05	.095	.052	.109	.054

Table 4-8. Reliability / dependability estimates for SSQ subscale D-studies

Type of Temperament	Number of Items	G-coefficient (reliability)	Φ (dependability)
EI	23	.71	.69
	36	.80	.78
PM	16	.57	.54
	28	.70	.66
	47	.80	.74
TF	10	.58	.56
	17	.70	.68
	28	.80	.78
OL	26	.73	.70
	23	.70	.67
	39	.80	.78

Note: EI = extrovert-introvert, PM = practical-imaginative, TF = thinking-feeling, OL = organized-flexible.

CHAPTER 5
DISCUSSION AND CONCLUSION

In this study, we have addressed the problem of analyzing an unbalanced data with a fixed facet (Shavelson & Webb, 1991; Brennan, 2001a)—the fact that a univariate analysis of such data is limited by the challenge from estimating variance components as well as relative less information derived from only one universe score. We have described another way, a multivariate analysis, of analyzing unbalanced data with a fixed facet. Multivariate analysis avoids the complexity of unbalanced design, and allows us to investigate and further decrease the measurement error at each subtest level and whole test. In addition, multivariate analysis provides information about disattenuated correlations between universe scores.

Based on multivariate analysis strategy, we used Student Style Questionnaire (SSQ) data to illustrate how to formulate and compute the variance components, covariance components, generalizability and phi coefficients (dependability index) at both micro-level (each profile) and macro-level (whole profiles). Our results lend support that multivariate analysis is a powerful method to investigate and design reliable observations for unbalance data with a fixed facet.

Discussion of the Results

SSQ measures four types of temperament: *extrovert-introvert* (OL), *practical-imaginative* (PM), *thinking-feeling* (TF) and *organized-flexible* (OL). SSQ includes 69 self-report bipolar items. Six SSQ items simultaneously measure two different types of temperament, which leads to 23 items measuring EI style, 16 items measuring PM style, 10 items measuring TF style and 26 items measuring OL style. Total 7902 students completed the questionnaire and eight random samples were produced for the convenience of data analysis.

Variance and covariance components for each sample size were obtained through ANOVA procedures. Results showed that the magnitudes of error variances are similar in eight random samples. Results from the G study using the mean variance and covariance components over eight samples indicated, in the profiles (EI, PM, TF and OL) of SSQ, that the variations due to persons (.31 to .69) and items (.44 to .68) are much smaller compared to the variation due to person by item interaction (3.68 to 4.89). The large variance from the interaction indicates the extent to which persons are ranked ordered differently by different items.

The Generalizability coefficients and dependability index for each SSQ profile are also obtained. OL has the largest G and phi coefficient (.73 and .70) while PM has the smallest G and phi coefficient (.57 and .54). By examining four profiles of SSQ, OL and EI have acceptable G and phi coefficients, but the G and phi coefficients for PM and TF are moderately low. Our subsequent D study results indicate that we can increase the G-coefficient (reliability) to .7 for PM and TF if we add 12 more items to PM and 7 more items to TF. Furthermore, if we expect the G-coefficient to be .8 for each SSQ subscale, we need 36 items for EI, 47 items for PM, 28 items for TF and 39 items for OL.

Multivariate analysis allows us to observe the disattenuated correlations between SSQ profiles. OL and TF have the highest correlations with PM (.36 and .32) while EI has the lowest correlation with PM and OL (-.02 and -.11), which indicates students preferring *organized*-type also prefer *thinking*-type and *practical*-type but not necessarily related with *extrovert*-type or *introvert*-type.

Based on a prior weights according to the importance of the number of items in each level of the fixed facet, the multivariate G and phi coefficients (.70 and .67) representing the overall SSQ

profiles are acceptable in this study. About 70% of the observed-score variance in all forced-choice items is due to the object of measurement (the person) in this study.

Limitations of This Study

In this study, there are several limitations. First, six out of 69 items measures two levels of the fixed facet (for example, item 41 measures OL as well as PM), so items are not purely nested within the profiles of SSQ. For the convenience of illustrating the method, we ignore this complexity of the design by treating the six items as different items to measure OL and PM, which caused the total number of items to be 75. Second, generalizability theory is most useful when many random facets are included in the research design. In this study, we only have one random face (items), besides object of measurement, which limits the current study in that G studies should usually involve as many as possible potential sources of variation. Third, the software SPSS and EXCEL instead of mGENOVA (Brennan, 2001b) were used to carry out the data analysis in some way leading to less precise calculation because of two or three digits rounding error.

Closing Remarks

The data that have been considered in this study involves unequal numbers of conditions nested within each level of a fixed facet, which is perfect for a multivariate design. Although a univariate method can also be used to analyze such data, estimating variance components would be much more difficult. Multivariate analysis is rather straightforward in this situation. When each level of the fixed facet contains an equal number of conditions, multivariate analysis constructs the variance components from each univariate analysis for each level of the fixed facet, and accounts for the covariance components that univariate analysis cannot.

LIST OF REFERENCES

Allen, M.J. & Yen, W.M. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, 22(4), 307-331.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, Vol.24, No. 4, 339-353

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

Brennan, R. L. (2001b). *mGENOVA(Version 2.1)* [Computer software and manual] Iowa City, IA: American College Testing, Inc.

Brennan, R. L.. Yin, P.. Kane, M. T. (2003). Methodology for Examining the Reliability of Group Mean Difference Scores. *Journal of Educational Measurement*. Vol. 40, No. 3, pp. 207-230.

Clauser, B. E., Harik, P. & Margolis M. J. (2006). A Multivariate Generalizability Analysis of Data from a Performance Assessment of Physicians' Clinical Skills. *Journal of Educational Measurement.* Vol. 43, No. 3, pp. 173-191.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

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

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, 16, 137-163.

Edward H. Haertel (2006). Reliability. *Educational measurement*, ch.3, 65-110.

Gao, X. & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, 14(2), 191-203.

Jung, C. G. (1971). *Psychological types.* (R.F.C. Hull, Revision of Trans. By H.G. Baynes). Princeton, NJ: Princeton University Press.

Kane, M. (1996). The Precision of Measurements. *Applied Measurement in Education*, 9(4), 355-379.

National Council on Measurement in Education and American Council on Education. (2006). *Educational Measurement*. 4th ed. Editor, Robert L. Brennan

Nubaum, Albert. & Aachen, Rwth. (1984). Multivariate Generalizability Theory in Educational Measurement: An Empirical Study. *Applied Psychological Measurement*. Vol. 8, No. 2, pp. 219-230.

Oakland, T., Glutting, J.J., & Horton, C. B. (1996). *Student Styles Questionnaire Manual*. San Antonio TX: Psychological Corporation.

Oakland, T. & Lu, Li. (2006). Temperament styles of children from the People's Republic of China and the United States. *School Psychology Internationally*, 27, 192-208.

Osterlind, S. J. (2005). Modern measurement: *Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson Prentice Hall.

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

Solano-Flores, G. & Li, M. (2006). The Use of Generalizability (G) Theory in the Testing of Linguistic Minorities. *Educational Measurement: Issue and Practice*. 25(1), 13-22.

Yin, Ping. (2005) A Multivariate Generalizability Analysis of the Multistate Bar Examination. *Educational and Psychological Measurement*, Vol. 65 No. 4, 668-686

Webb, N. M. & Shavelson, R. J. (1981). Multivariate Generalizability of General Educational Development Ratings. *Journal of Educational Measurement*. V18. NO. 1, 13-22.

BIOGRAPHICAL SKETCH

Youzhen Zuo received the Bachelor of Art in English in 1992 from Wuhan University in China. She worked as an English teacher for over seven years in Fanchuan Middle School. In 2004, she was admitted to Ohio State University as a graduate student at the School of Educational Policy and Leadership. In 2005, she transferred as a graduate student to the Department of Educational Psychology at University of Florida.