Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 10

5-1-2003

Incorporating Sampling Weights Into The Generalizability Theory For Large-Scale Analyses

Christopher W. T. Chiu
Law School Admissions Council, cchiu@lsac.org

Ronald S. Fesco
National Science Foundation

Incorporating Sampling Weights Into The Generalizability Theory For Large-Scale Analyses

Cover Page Footnote
The authors thank Robert Brennan, Neil Timm, and Loan Tran for their suggestions and comments.

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No. 1, 108-127

1538 – 9472/03/$30.00

Incorporating Sampling Weights Into The Generalizability Theory For Large-Scale Analyses

Christopher W.T. Chiu
Law School Admission Council

Ronald S. Fecso
National Science Foundation

Large scale studies frequently use complex sampling procedures, disproportionate sampling weights, and adjustment techniques to account for potential bias due to nonresponses and to ensure that results from the sample can be generalized to a larger population. Survey researchers are concerned about measurement error and the use of weights in developing models. Consequently, multiple weighting factors are used and these weighting factors are manifested as a final survey (composite) weight available for analysis. We developed a method to incorporate an external weighting factor like this for analyses of measurement errors in the theory of generalizability to provide researchers with a tool to evaluate the measurement error components of survey quality and undesirable error components of large-scale assessment programs such as national and state assessments.

Key words: Generalizability theory, large-scale performance assessment, rater reliability, sampling, Survey of Doctorate Recipients (SDR), variance component, weighting

Introduction

The focus of this research is to illustrate how to incorporate weights in the framework of generalizability theory (Brennan, 1992a; Cronbach, Gleser, Nanda, and Rajaratnam, 1972; and Shavelson and Webb, 1991) when it is applied to large-scale studies such as national surveys and educational assessments.

Chris W. T. Chiu is a Research Scientist, Psychometrics Group, Law School Admission Council (LSAC), 661 Penn Street, Newtown, PA 18940. Email: cchiu@lsac.org, Ronald S. Fecso is Chief Statistician, National Science Foundation (NSF). This work was partially supported by the American Statistical Association (ASA) through a grant from NSF, Division of Science Resources Statistics (grant number: SRS-0004192). A portion of the research was conducted while Chris Chiu was a professor at the University of Pittsburgh. The authors thank Robert Brennan, Neil Timm, and Loan Tran for their suggestions and comments. Information in this article represents the opinions of the authors and is not NSF, the ASA, the University of Pittsburgh, and the LSAC.

This research is important because educational researchers need to determine variance components and reliability coefficients to accurately reflect measurement errors in statewide or nationwide assessment programs, which often test only a sample of students for accountability purposes. Generalizability theory is a well-known method in educational and psychological research, but today, no one has examined the effect of sample survey data on the method. In addition, survey researchers can use such knowledge to understand, monitor, and improve survey quality. If a weighting scheme was used but researchers ignored the weights in generalizability studies (G studies), as is often the case with such a model, the estimated errors will be biased (Rosenbaum, 1987). In addition, the standard error of the variance component estimates will be inappropriate.

A very popular model in generalizability theory is the two-facet crossed model, which is frequently used in monitoring measurement errors (e.g., Brennan et al., 1995, Brennan, 2000b; Chiu and Wolfe, 2002; Lane et al., 1996) when human judgments are involved. The model can partition error variances into specific sources so that researchers can determine which error source(s) is/ are most in need for reduction. For example, one

can determine the score consistency in high-stake examinations where test-takers respond to a set of test questions scored by a group of raters (i.e., a *person x item x rater* two-facet model). Alternatively, one can use a two-facet crossed model (i.e., *respondent x item x coding method*) to determine the coding consistency in survey analysis where survey responses are coded using different schemes (e.g., self-report versus objectively coded responses).

Despite the common applications of the generalizability theory in survey studies (Adam and Ujwal, 1999; Johnson and Bell, 1985; Shipper, et al., 1986), we did not find references discussing how one could incorporate weights into G studies — we searched monographs on G theory (Brennan, 1992a; Brennan, 2001b; Chiu, 2001; Cronbach, et. al., 1972; Fyans, 1983; Shavelson and Webb, 1991) and on variance estimations (Rao, 1997; and Wolter, 1985) using the five major modes of searching: footnote chasing, consultation, searches in subject indices, browsing, and citation searchers (White, 1994). Also, we contacted experts in G theory (Brennan, 2001b; Cronbach, 2000) and searched journal articles and electronic databases (PSYINFO, 1887–2001; ERIC, 1966-2001; MEDLINE, 1966-2001; JSTOR, 1887-1996; Sociological Abstracts, 1963-2001).

In the current study, we first reviewed the purposes and importance of survey weights followed by a summary of the traditional variance component estimation procedures. Second, we discussed the concepts and essential steps of a new weighting method in G studies (i.e., the Chiu-Fecso G-method, denoted CFG hereafter). Specifically, we used two examples to illustrate the method. The first example was a hypothetical dataset with a context in educational assessment and the other was an operational dataset from a large-scale survey used for research on science and engineering education. (The Survey of Doctorate Recipients is a longitudinal survey administered by the Division of Science Resources Statistics (SRS) at the National Science Foundation (NSF). Details of the survey can be found in the homepage of SRS: http://www.nsf.gov/sbe/srs). We intentionally used a simple case in the first example to demonstrate the computational procedures of the new method. The example was simple enough for hand calculation. The second example, based on an operational dataset from a national study, was used to show the capacity of the method for a real data set. Given the wide applications of the two-facet crossed model, we focus our discussions on the two-facet model throughout the manuscript.

Basic Concepts of G Theory and Weighting

An extension of the Classical Test Theory (Crocker and Algina, 1986) and the Analysis of Variance (ANOVA) methods, G theory has been applied to examine the reliability and validity of measurement procedures in educational assessments, psychological measurement, program evaluations, and survey analysis. As Shavelson and Webb (1991) stated:

> "The strength of G theory is that multiple sources of error in a measurement can be estimated separately in a single analysis. Consequently, in a manner similar to the way the Spearman-Brown 'prophecy formula' is used to forecast reliability as a function of test length in classical test theory, G theory enables the decision maker to determine how many occasions, test forms, and administrators are needed to obtain dependable scores. In the process, G theory provides a summary coefficient reflecting the level of dependability, a generalizability coefficient that is analogous to classical test theory's reliability coefficient." (p. 2)

Brennan (1992a, 1992b, and 2000a) and Shavelosn and Webb (1991) provided a succinct treatment of the essential features of G theory. Chiu (1999a, 2001) developed a subdividing method to estimate variance components in large-scale performance assessments with missing observations. Brennan (2000a) discussed the misconceptions about the theory. Brennan and Johnson (1995) and Cronbach, Linn, Brennan, and Haertel (1997) covered basic concepts in G theory. Brennan (1997) and Shavelson and Webb (1981) summarized the history of the G theory. Despite the popularity of G theory, all of the

aforementioned studies assumed that simple random sampling was used.

Traditionally, G theory assumes less than or equal to simple random sampling (Bell, 1985; Brennan, 1992a; Cronbach et al., 1972), only that every person has the same probability of being sampled from a population or, that every element is assigned a unit weight. Such an assumption is not viable in national studies where complex sampling procedures (e.g., disproportionate sampling of smaller demographic groups) are used. To create representative estimates in such cases, variable probabilities of selection or variable weights are needed.

Another purpose of weighting is to adjust for the effects of non-respondents (Kish, 1995; Lee, Forthofer, and Lorimer, 1989; and Sarndal, 1980). Bailar, Bailey, and Corby (1978) summarized the purposes and compared some adjustment and weighting procedures (e.g., reweighting, substitution, regression) that were actually used at the US Bureau of the Census, for survey data. The National Science Foundation provided a concise summary of using survey weights, for the Survey of Doctorate Recipients (SDR) — a longitudinal panel survey of individuals who have received their doctorates mainly in the sciences or engineering fields (the data of this survey is used as an example in subsequent sections):

> Sampling weights were defined as the reciprocal of the probability of selection for each sampled units, and the weights were adjusted by using weighting class or poststratification adjustment procedures. The final adjusted sampling weights become the analysis weights [also called Final Survey Weights], which have been added to each individual's record in the survey database. (Author, 2002)

Instead of making available multiple weights to researchers, survey developers create a single composite weight also called the final survey weight (e.g., in the Survey of Doctorate Recipients) for analysis. Designed as a proxy for all the weighting factors in the survey, the Final Survey Weights may be the only weighting information available in the survey data. In this paper, we first derived the methodological adjustments to incorporate such a composite weight on G theory estimation. We then applied the methodology in the context of a large-scale survey to examine the impact of the methodological change and substantively the occupational stability in the engineering profession of the United States. The methodology developed here can be used directly in any crossed design with two facets. The three principles of the weighting method discussed in this paper, however, can be used for other designs with any number of facets. However, our intention is to focus on a two facet crossed design, which has a variety of applications in measurement.

Methodology

Detecting Measurement Errors and Estimating Variance Components

Many have contributed to the methods in monitoring measurement errors and in estimating variance components. In the survey research context, Biemer and Fecso (1995), Rao and Sitter (1997), and Reiser, Fecso, and Chua (1992) discussed methods to characterize measurement errors. In the statistics and educational assessment context, Brennan (1992a), Chiu (1999a, 1999b), Chiu and Wolfe (1997), Corbeil and Searle (1976), Millman and Glass (1967), and Searle, Casella, and McCulloch (1992) among others, provided in-depth discussions on variance component estimation methods. Brennan (1992a) offered an extensive treatment on the topic geared toward generalizability theory. Also, he used synthetic datasets to illustrate the computational steps for variance component estimations. Instead of repeating the details, we summarized the general procedures below and used the summary as building blocks to develop a weighted variance component method based on G theory discussed in the subsequent sections.

In G theory, variance component estimates can be obtained by solving a set of Expected Mean Square (EMS) equations (Brennan, 1992a, chapter 2 and 3; appendices A through B) relating the variance components and mean squares. In the sections that follow, we used a fully crossed two-

faceted design (Brennan, 1992a) as an example. Unless stated otherwise, the universe of admissible observations contains person (p), item (i), and rater (r). The EMS equations can be expressed in the following matrix formula,

$$\hat{\mathbf{s}}^2 = \mathbf{C}\,\hat{\mathbf{a}}^2 \qquad (1)$$

where $\mathbf{C}$ is an f x f upper-triangular matrix of coefficients for the variance components estimated, and f = 1,2, …, 7 represent the seven variance component estimates in a two faceted design. The column vector $\hat{\mathbf{a}}^2$ is a set of mean squares for the effects observed in the data (Brennan, 1992a). One can also explicitly write out the elements in $\mathbf{C}$ and $\hat{\mathbf{a}}^2$ as follows.

$$\begin{bmatrix} \hat{s}^2_p \\ \hat{s}^2_i \\ \hat{s}^2_r \\ \hat{s}^2_{pi} \\ \hat{s}^2_{pr} \\ \hat{s}^2_{ir} \\ \hat{s}^2_{pir} \end{bmatrix} = \begin{bmatrix} (n_i n_r)^{-1} & 0 & 0 & (n_r)^{-1} & (n_i)^{-1} & 0 & 1 \\ 0 & (n_p n_r)^{-1} & 0 & (n_r)^{-1} & 0 & (n_p)^{-1} & 1 \\ 0 & 0 & (n_p n_i)^{-1} & 0 & (n_i)^{-1} & (n_p)^{-1} & 1 \\ 0 & 0 & 0 & (n_r)^{-1} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & (n_i)^{-1} & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & (n_p)^{-1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} MS_p \\ MS_i \\ MS_r \\ MS_{pi} \\ MS_{pr} \\ MS_{ir} \\ MS_{pir} \end{bmatrix} \qquad (2)$$

The mean squares vector $\hat{\mathbf{a}}^2$, in the above, can be estimated by dividing the set of "sum of squared means" by their corresponding degrees of freedom (Brennan, 1992a, p. 36). We represented such computations using Equation (3), whose elements are explicitly shown in Equation (4).

$$\hat{\mathbf{a}}^2 = \mathbf{D}\,\mathbf{t} \qquad (3)$$

$$\hat{\mathbf{a}}^2 = \begin{bmatrix} MS_p \\ MS_i \\ MS_r \\ MS_{pi} \\ MS_{pr} \\ MS_{ir} \\ MS_{pir} \end{bmatrix} = \begin{bmatrix} (n_p-1)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & -(n_p-1)^{-1} \\ 0 & (n_i-1)^{-1} & 0 & 0 & 0 & 0 & 0 & -(n_i-1)^{-1} \\ 0 & 0 & (n_r-1)^{-1} & 0 & 0 & 0 & 0 & -(n_r-1)^{-1} \\ -(n_p-1)^{-1}\bullet(n_i-1)^{-1} & -(n_p-1)^{-1}\bullet(n_i-1)^{-1} & 0 & (n_p-1)^{-1}\bullet(n_i-1)^{-1} & 0 & 0 & 0 & (n_p-1)^{-1}\bullet(n_i-1)^{-1} \\ -(n_p-1)^{-1}\bullet(n_r-1)^{-1} & 0 & -(n_p-1)^{-1}\bullet(n_r-1)^{-1} & 0 & (n_p-1)^{-1}\bullet(n_r-1)^{-1} & 0 & 0 & (n_p-1)^{-1}\bullet(n_r-1)^{-1} \\ 0 & -(n_i-1)^{-1}\bullet(n_r-1)^{-1} & -(n_i-1)^{-1}\bullet(n_r-1)^{-1} & 0 & 0 & (n_i-1)^{-1}\bullet(n_r-1)^{-1} & 0 & (n_i-1)^{-1}\bullet(n_r-1)^{-1} \\ (n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & (n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & (n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & -(n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & -(n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & -(n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & -(n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} & -(n_p-1)^{-1}\bullet(n_i-1)^{-1}\bullet(n_r-1)^{-1} \end{bmatrix} \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{pir,e} \\ T_m \end{bmatrix} \qquad (4)$$

The elements of the **D** matrix in equations (3) and (4) are the sample sizes (n_p, n_i, n_r) involved in the seven variance components of the two faceted crossed design. The "sum of squared mean" denoted T_f is computed for each facet and for the grand mean, such that $\mathbf{t} = [T_1, ..., T_f]'$. The rightmost side of equations (3) and (4), $\mathbf{t}$, can be computed by summing individual scores, taking the average, squaring the mean, and multiplying the squared mean by the number of levels in the facet(s) other than the facet for which the sum of squared mean is computed. See equation (5).

$$\mathbf{t} = \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{pire} \\ T_m \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p \bar{x}^2_{p\cdot\cdot} \\ n_p n_r \sum_i \bar{x}^2_{\cdot i\cdot} \\ n_p n_i \sum_r \bar{x}^2_{\cdot\cdot r} \\ n_r \sum_p \sum_i \bar{x}^2_{pi\cdot} \\ n_i \sum_p \sum_r \bar{x}^2_{pr\cdot} \\ n_p \sum_i \sum_r \bar{x}^2_{\cdot ir} \\ \sum_p \sum_i \sum_r x^2_{pir} \\ n_p n_i n_r \bar{x}^2 \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p ((\frac{1}{n_i n_r} \sum_i \sum_r x_{pir})^2) \\ n_p n_r \sum_i (\sum_p (\frac{1}{n_r} \sum_r x_{pir}))^2 \\ n_p n_i \sum_r (\sum_p (\frac{1}{n_i} \sum_i x_{pir}))^2 \\ n_r \sum_p \sum_i (\frac{1}{n_r} \sum_r x_{pir})^2 \\ n_i \sum_p \sum_r (\frac{1}{n_i} \sum_i x_{pir})^2 \\ n_p \sum_i \sum_r \sum_p (x_{pir})^2 \\ \sum_p \sum_i \sum_r x^2_{pir} \\ n_p n_i n_r (\frac{1}{n_p n_i n_r} \sum_p \sum_i \sum_r x_{pir})^2 \end{bmatrix} \quad (5)$$

Conceptual Framework of the Chiu-Fecso G-Method

One limitation of the traditional method is that it assumes that every person carries the same weight in an analysis. This assumption is often violated in sample surveys where persons typically receives a different weight as a result of complex sampling and valid response adjustments discussed earlier (See Basic Concepts of G Theory and Weighting). The Chiu-Fecso method enables such a weight (a composite weight supplied to analysts by survey developers and statisticians) to be incorporated in generalizability studies. See Equation (5) for the "sum of squared mean" shown in the $\mathbf{t}$ vector. Prior to a thorough treatment in computing the weighed sum of squared means, we introduced three fundamental principles used in the Chiu-Fecso G-method.

Multiplication Principle

The summations in Equation (5) simply add up individual scores, assuming that each score occurs once in the data. For example, the total of a set of scores {2, 1, 3, 4} is obtained by 1•2 + 1•1 + 1•3 + 1•4 = 10. This approach, assuming that each score received a unit weight, is used in the traditional framework of G theory (Brennan, 1992a, 1992b), discussed in the previous section. The Chiu-Fecso approach relaxed such assumption by allowing each score to have a different weight. This difference is critical when incorporating survey weights for computing the "sum of squared means" because the idea of using survey weights is equivalent to replicating an observed value by the number of times specified in the weights. Rosenbaum (1987) called such weighting approach "direct adjustment." He pointed out that direct adjustment has two attractive properties: (a) it does not require explicit modeling of the stratification in the sampling design and (b) it produces parallel adjustments in the original statistical procedures so that only little modifications are needed in adapting the original procedures. Consistent with Rosenbaum (1987), Lee, Forthofer, and Lorimor (1989) advocated the use of weights, which they called the weights "expansion weights," to compute unbiased estimates for means and sums. However, they did not develop a method for variance components. This limitation motivates the current study. To begin, we review the expansion weights. First, assume that the first two scores {2, 1} in the previous example came from a minority group, and each received a composite weight of 49. Further assume that the last two scores came from a majority group and thus received a unit composite weight. The total became 49•2 + 49•1 + 1•3 + 1•4 = 154. In the following two sections, we modified the "expansion weight" to obtain the adjusted degrees of freedom (using the Adjustment Principle) and the weighted mean (using the Relative Weighting Principle). These two quantities serve as the building blocks for the weighted variance components discussed in the subsequent section (Computational Equation of the Chiu-Fecso Method).

Adjustment Principle

The goal of inferential statistics is to determine the extent to which we can infer the results from a sample to a target population. A critical factor in making correct inferences is to determine the correct degrees of freedom reflecting the sample size. In the previous example, a sample size of 4 was collected and each person received a weight assigned by survey developers, statisticians, or policy makers. As shown earlier, if we were to apply the multiplication principle directly, we would obtain a total of 154 (49•2 + 49•1 + 1•3 + 1•4 = 154). However, this approach is problematic because it assumes that a sample of 100 was collected (49+49+1+1). Put differently, this approach erroneously expanded the degrees of freedom. To correct for this problem, we use an adjustment principle so that the weights reflect the actual sample size (n = 4) and also the correct degrees of freedom. Such adjustment is accomplished through dividing each weight in the vector of weight $\mathbf{w}$ = [49 49 1 1] by the mean of the weights (Σw_p/n). After the adjustment, the "adjusted expansion weights" became $\mathbf{w}$ / (Σ w_p / n) = [49 49 1 1] / 25 = [1.96 1.96 0.04 0.04]. Note that the total of the adjusted expansion weights matches the sample size (n = 4) and the ratio between the first and third cases remains 49 to 1. In general, the ratios among all the cases remain unchanged.

Relative Weighting Principle

One way to obtain the weighted mean for a set of values is to add up all the weighted scores in a set and then divided the total by the total weight or the number of scores in the set, ($\Sigma wx/\Sigma w$). An alternative is to multiply each unique value of a set of scores by its relative frequency and then add up the products (i.e., $\Sigma f(x)\bullet x$). For instance, the weighted average of the previous example is 0.49•2 + 0.49•1 + 0.01•3 + 0.01•4 = 1.54, where 0.49 was obtained by dividing the sampling weight for the first case by the total weight of the four cases (i.e., 49 / 100). Hereafter we referred to $f(x)$ as the relative frequency.

With the multiplication principle, the adjustment principle, and the relative weighting principle, we have computed the adjusted total, adjusted degrees of freedom, and adjusted means in the above sections. Next we introduce the CFG method to analytically compute the weighted variance component estimates.

Computational Equation of the Chiu-Fecso Method

An assumption and three steps are involved in our modification of the G theory. We assume that a set of composite weights is given and stored in a row vector $\mathbf{w}$. With this set of weights, we first compute the adjusted expansion weights (using the adjustment principle). Second, we compute the relative weights based on the adjusted expansion weights (using the relative weighting principle). Third, we apply two decision rules to determine when and how to use the two sets of weights obtained in steps 1 and 2.

Step 1: Compute Adjusted Expansion Weights

In general, a row vector of the adjusted expansion weights ($\mathbf{w}_p$) is obtained by dividing each of the weights in $\mathbf{w}$ by the mean of all the weights. That is, $\mathbf{w}_p = [w_1 \ w_2 \ w_3 \ \ldots \ w_p] / (\Sigma w/n)$.

Step 2: Compute Relative Weights

The relative weights, denoted $\mathbf{w}_{f(p)}$, are obtained by dividing each of the adjusted expansion weights above by the sum of these weights. That is, $\mathbf{w}_{f(p)} = [w_{p_1} \ w_{p_2} \ w_{p_3} \ \ldots \ w_{p_p}] / (\Sigma w_p)$. Since the sum of all the adjusted expansion weight equals to the sample size, an alternative is: $\mathbf{w}_{f(p)} = [w_{p_1} \ w_{p_2} \ w_{p_3} \ \ldots \ w_{p_p}] / n$.

Step 3: Apply Decision Rules

Rule #1: When finding the weighted sum in a facet of interest, we pre-multiply the adjusted expansion weighting vector ($\mathbf{w}_p$, a row vector) to a set

of scores (**s**, a column vector), resulting in $\mathbf{w}_p \bullet \mathbf{s}$.

Rule #2: When finding the weighted average score in the facet of interest, we pre-multiply the vector of relative weights to the column vector of scores (i.e., $\mathbf{w}_{f(p)} \bullet \mathbf{s}$).

How do we apply the two decision rules to the theory of generalizability? We replace all $\sum_p$ in Equation (5) with $\sum_p w_p$ when the facet of interest involves the weighting facet (in this case, the Object of Measurement, person); otherwise, we replace $\sum_p$ in Equation (5) with $\sum_p w_{f(p)}$. For example, the first entry in **t** of Equation (5) is the Object of Measurement (p), which is also the weighting facet, so we insert w_p to $\sum_p$, resulting $\sum_p w_p$. In the second entry of **t** of Equation (5), the facet of interest involves item (i) and does not involve the weighting facet (p), so we replace $\sum_p$ with $\sum_p w_{f(p)}$. By the same token, we apply the same rule to the remaining entries in **t** of Equation (5). Consequently, we have Equation (6). We highlighted w_p in circle and $w_{f(p)}$ in square to show where to insert the weights.

$$\mathbf{t}^{(w)} = \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{pir,e} \\ T_m \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p \bar{x}^2_{p..} \\ n_p n_r \sum_i \bar{x}^2_{i..} \\ n_p n_i \sum_r \bar{x}^2_{r..} \\ n_r \sum_p \sum_i \bar{x}^2_{pi.} \\ n_i \sum_p \sum_r \bar{x}^2_{pr.} \\ n_p \sum_i \sum_r \bar{x}^2_{ir.} \\ \sum_p \sum_i \sum_r x^2_{pir} \\ n_p n_i n_r \bar{x}^2 \end{bmatrix} = \begin{bmatrix} n_i n_r \sum_p w_p \left(\frac{1}{n_i n_r} \sum_i \sum_r x_{pir}\right)^2 \\ n_p n_r \sum_i \left(\sum_p w_{f(p)} \left(\frac{1}{n_r} \sum_r x_{pir}\right)\right)^2 \\ n_p n_i \sum_r \left(\sum_p w_{f(p)} \left(\frac{1}{n_i} \sum_i x_{pir}\right)\right)^2 \\ n_r \sum_p w_p \left(\sum_i \left(\frac{1}{n_r} \sum_r x_{pir}\right)^2\right) \\ n_i \sum_p w_p \left(\sum_r \left(\frac{1}{n_i} \sum_i x_{pir}\right)^2\right) \\ n_p \sum_i \sum_r \sum_p \left(w_{f(p)} x_{pir}\right)^2 \\ \sum_p w_p \sum_i \sum_r x^2_{pir} \\ n_p n_i n_r \left(\frac{1}{n_p n_i n_r} \sum_p w_p \sum_i \sum_r x_{pir}\right)^2 \end{bmatrix} \quad (6)$$

where w_p is the adjusted expansion weight for person *p* and $w_{f(p)}$ is the relative weight for person *p*.

With the updated "sum of mean scores" in Equation (6), we obtained the weighted variance component estimates using the following steps. First, compute the weighted "sum of mean scores" vector ($\mathbf{t}^{(w)}$) as shown in Equation (6). Second, substitute $\mathbf{t}^{(w)}$ back to Equation (4) to obtain the updated Mean Squares $[\hat{\mathbf{a}}^2]^{(w)}$, which in turn is substituted back to equation (2) to obtain weighted variance component estimates $[\hat{\mathbf{s}}^2]^{(w)}$. In summary, we estimate the weighted variance component estimates using:

$$\left[\underset{7x1}{\hat{\mathbf{s}}^2}\right]^{(w)} = \underset{7x7}{\mathbf{C}}\ \underset{7x8}{\mathbf{D}}\ \underset{8x1}{\mathbf{t}^{(w)}} \qquad (7)$$

The standard error of the weighted variance components can be obtained by substituting the weighted means squares $MS_j^{(w)}$, their coefficients c_j, and degrees of freedom df_j into Equation (8). Brennan (1992a) and Chiu (1999a) provided an in-depth discussion for the unweighted standard error equations. Chiu (2001, p. 127, Equations 34 through 40) expressed the standard errors in terms of variance components and the number of levels in each facet. Brennan (1992a, p. 101, equation 6.2.1) provided the general form of the equation. We modified the general equation to incorporate the composite weights as follows:

$$[\mathrm{SE}(\hat{s}_f^2)]^{(w)} = \sqrt{\sum_j \frac{2(c_j MS_j^{(w)})^2}{df_j + 2}} \qquad (8)$$

One cautious note to Equation (8) is the distinction between the subscripts f and j. The former denotes the f^{th} variance component and the latter denotes the j^{th} Mean Square term for the f^{th} variance component. As shown in Equation (2), each variance component estimate involves a different number of Mean Square terms and for this reason, J, the total number of mean square terms varies for each variance component estimate. For simplicity and consistency with the G theory literature, we use a single subscript notation j as opposed to the double subscript notation j_f, although they are interchangeable in this context.

Results

Validation of the Weighted Method

Being able to incorporate weights in generalizability studies are particularly important when the weights differ greatly among the samples. We used a published data set with 10 hypothetical cases and purposely assigned highly disproportionate weights to the data set (one case received a weight of 10 while the rest received a unit weight). As a result, the ratio of the weighted and unweighted variance component estimates was between 0.3459 and 2.9865, for the seven components, indicating that the weighted estimates could be almost three times larger or three times lower than the unweighted estimates (See Appendix B). Such a result reminds researchers that weighted estimates could be different from their unweighted counterparts when extreme values appear in the weights. The extent to which the two types of estimates would become drastically different depends on the weighting scheme provided in the survey.

We purposely chose an extreme example to contrast the weighted and unweighted results. Such an example is realistic because when applying a two-facet model where test items or tasks are involved, researchers may desire to explore the effect of assigning a much larger weight to one important item — a 300 word essay requiring 45 minutes of testing time may be weighted as much as 10 times more than a multiple-choice question requiring lower than two minutes of testing time.

The aforementioned example (discussed fully in Appendix B) also served as a benchmark comparison between the Chiu-Fecso method and the traditional unweighted method (Brennan, 1992a). Appendix B shows that the unweighted method was a special case of the weighted method because when the weights were set to unity, the CFG method yielded identical variance component estimates to the traditional method.

Example 1: Performance Assessment

Performance assessment has been popular in the recent decades (Bejar and Braun, 1999; Bennett and Sebrechts, 1996; Braun, Bennett, Frye, and Soloway, 1990; Brennan, 2000b; Chiu, 2001; Clauser, 2000). Many educational and professional testing programs employ constructed-response items to assess performance (e.g. the National Assessment of Educational Progress, the Texas Assessments of Academic Skills, and the United States Medical Licensing Examination). Generalizability analysis is one of the popular techniques to examine the quality of test scores and it can provide guidance regarding the potential to reduce measurement error (Brennan, 2000b; Clauser, 2000). Of the many models in G theory, the two-facet crossed model (Brennan, 2000; Chiu, 2001) is frequently used. Utilizing a two-faceted

model, the following hypothetical data set (3 items *x* 2 raters) demonstrates the computational procedures of the Chiu-Fecso method. As shown in the data matrix **X**, each of the four persons has six scores arranged in a row. Columns one through three represent the scores on the three items judged by the first rater; Columns four through six represent the scores on the same three items judged by the second rater. The gap between the third and forth columns is intended to visually separate the scores for the two raters.

$$\mathbf{X} = \begin{bmatrix} 1\ 1\ 1 & 0\ 0\ 0 \\ 1\ 1\ 1 & 1\ 0\ 1 \\ 1\ 0\ 0 & 0\ 1\ 1 \\ 0\ 0\ 0 & 0\ 0\ 1 \end{bmatrix} \qquad (9)$$

Assume that a final survey weight is derived by survey developers and it is the only weighting information available in the data given to the analyst. Further assume that the weights for the four persons are stored in a row vector [2 3 4 1] which is given to the analyst. We then obtained the adjusted expansion weights and relative weights as follows.

$\mathbf{w_p}$ = [0.8 1.2 1.6 0.4] = [2 3 4 1] / ((2 + 3 + 4 +1) / 4) and

$\mathbf{w_{f(p)}}$ = [0.2 0.3 0.4 0.1] = [0.8 1.2 1.6 0.4] / ((0.8 + 1.2 + 1.6 + 0.4)).

With the $\mathbf{w_p}$ and $\mathbf{w_{f(p)}}$ computed, we used Equation (6) to obtain $\mathbf{t}^{(w)}$ as shown below (see Appendix A for the step-by-step illustrations).

$$\mathbf{t}^{(w)} = \begin{bmatrix} T_p \\ T_i \\ T_r \\ T_{pi} \\ T_{pr} \\ T_{ir} \\ T_{pire} \\ T_m \end{bmatrix}^{(w)} = \begin{bmatrix} 6.8667 \\ 6.6200 \\ 6.4133 \\ 7.4000 \\ 8.4000 \\ 7.0000 \\ 12.400 \\ 6.4067 \end{bmatrix} = \begin{bmatrix} 3 \text{ x } 2 \text{ x } 1.1445 \\ 4 \text{ x } 2 \text{ x } 0.8275 \\ 4 \text{ x } 3 \text{ x } 0.5344 \\ 2 \text{ x } 3.7000 \\ 3 \text{ x } 2.8000 \\ 4 \text{ x } 1.7500 \\ 12.4000 \\ 4 \text{ x } 3 \text{ x } 2 \text{ x } 0.2669 \end{bmatrix} \qquad (10)$$

By using $n_p = 4$, $n_j = 3$, and $n_r = 2$, and equation (4), we post-multiplied $\mathbf{t}^{(w)}$ to $\mathbf{D}$. The product became the weighted mean square vector $[\mathbf{a}^2]^{(w)}$. See equation (11)

$$[\mathbf{a}^2]^{(w)} = \begin{bmatrix} MS_p \\ MS_i \\ MS_r \\ MS_{pi} \\ MS_{pr} \\ MS_{ir} \\ MS_{pir} \end{bmatrix}^{(w)} = \begin{bmatrix} 0.1533 \\ 0.1067 \\ 0.0067 \\ 0.0533 \\ 0.5089 \\ 0.1867 \\ 0.5156 \end{bmatrix}$$

$$= \begin{bmatrix} (3)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 & -(3)^{-1} \\ 0 & (2)^{-1} & 0 & 0 & 0 & 0 & 0 & -(2)^{-1} \\ 0 & 0 & (1)^{-1} & 0 & 0 & 0 & 0 & -(1)^{-1} \\ -(3)^{-1} \bullet (2)^{-1} & -(3)^{-1} \bullet (2)^{-1} & 0 & (3)^{-1} \bullet (2)^{-1} & 0 & 0 & 0 & (3)^{-1} \bullet (2)^{-1} \\ -(3)^{-1} \bullet (1)^{-1} & 0 & -(3)^{-1} \bullet (1)^{-1} & 0 & (3)^{-1} \bullet (1)^{-1} & 0 & 0 & (3)^{-1} \bullet (1)^{-1} \\ 0 & -(3)^{-1} \bullet (1)^{-1} & -(2)^{-1} \bullet (1)^{-1} & 0 & 0 & (2)^{-1} \bullet (1)^{-1} & 0 & (2)^{-1} \bullet (1)^{-1} \\ (3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & (3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & (3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & -(3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & -(3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & -(3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & (3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} & -(3)^{-1} \bullet (2)^{-1} \bullet (1)^{-1} \end{bmatrix} \begin{bmatrix} 6.8667 \\ 6.6200 \\ 6.4133 \\ 7.4000 \\ 8.4000 \\ 7.0000 \\ 12.400 \\ 6.4067 \end{bmatrix} \qquad (11)$$

Next, we post-multiplied the mean square vector $[\mathbf{a}^2]^{(w)}$ to the $\mathbf{C}$ matrix to obtain the variance component estimates. See equation (12). Note that negative variance component estimates occurred in the hypothetical example because we used a randomly generated hypothetical data set, which had only a small sample ($n_p = 4$). Also, for simplicity, no distribution assumptions were specified in generating the data. In practice, one may not obtain negative estimates. Cronbach et. al. (1972) and Brennan (1992a) discussed the causes of negative variance components and developed methods to avoid negative variance component estimates. Those methods include Algorithm 2 (Brennan, 1992a) and Bayesian procedures (see Box and Tiao, 1973; Searle, et al., 1992).

$$\begin{bmatrix} \hat{\sigma}^2_p \\ \hat{\sigma}^2_i \\ \hat{\sigma}^2_r \\ \hat{\sigma}^2_{pi} \\ \hat{\sigma}^2_{pr} \\ \hat{\sigma}^2_{ir} \\ \hat{\sigma}^2_{pir} \end{bmatrix}^{(w)} = \begin{bmatrix} 0.0178 \\ 0.0478 \\ (-0.0144) \\ (-0.2311) \\ (-0.0022) \\ (-0.0822) \\ 0.5156 \end{bmatrix}$$

$$= \begin{bmatrix} (3\bullet 2)^{-1} & 0 & 0 & (2)^{-1} & (3)^{-1} & 0 & 1 \\ 0 & (4\bullet 1)^{-1} & 0 & (2)^{-1} & 0 & (4)^{-1} & 1 \\ 0 & 0 & (4\bullet 3)^{-1} & 0 & (3)^{-1} & (4)^{-1} & 1 \\ 0 & 0 & 0 & (2)^{-1} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & (3)^{-1} & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & (4)^{-1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.1533 \\ 0.1067 \\ 0.0067 \\ 0.0533 \\ 0.5089 \\ 0.1867 \\ 0.5156 \end{bmatrix} \quad (12)$$

Example 2: Large-Scale Survey Analysis

A panel sample of 2388 Engineers was obtained from a longitudinal survey for doctorate recipients. The survey was administered biennially. All survey respondents in the selected sample (a) were under the age 76, in 1999; (b) received at least one research doctorate in Science or Engineering from a U.S. institution in or prior to 1990; (c) were residing in the States on April 15 in four survey years analyzed in the current study (1993, 95, 97, and 99); and (d) were employed in the Engineering profession for at least one of the four aforementioned survey years. The panel of 2388 Engineers represented a population of approximately 50832 Engineers in the U.S. Engineers were broadly defined as those employed in professions such as Aerospace Engineering, Chemical Engineering, Civil and Architectural Engineering, Electrical, Electronic, Computer and Communications Engineering, Industrial Engineering, Mechanical Engineering, Postsecondary Engineering Teaching, and other Engineering fields. Using their age in 1999, the 2388 Engineers with Ph.D degrees can be divided into the following age groups.

Age Groups	Below 30	35-39	40-44	45-49	50-54
Sample Size	3	202	439	400	440
Age Groups	55-59	60-64	65-69	Above 70	
Sample Size	392	256	140	116	

Respondents were given a list of 126 job codes and were asked to choose the most appropriate title for their principal jobs (i.e., self-reported job codes). In addition, the respondents also reported their employment history and background information (e.g., sector of employment, work activities, number of people supervised directly). Such information was used to derive a second measure of occupational title, which was called the "best codes" of occupational titles. The best codes were derived using employment history, job activities, and such. Comprehensive discussions of the best coding process can be found in Hardy and Eisenhower (1994), McGuinness (1997), Rak, Chen, and Gray (1997).

Due to complex sampling and adjustment of nonresponse rate, respondents were selected with a different probability and thus a weighting scheme was used to ensure the representativeness of the sample. The average weight for Engineers was 21.29 (SD = 9.71; median = 22.98; minimum = 1.05; maximum = 46.72).

We conducted a generalizability study with a crossed design (G study, Brennan, 1992a; 1992b) to measure occupational changes. Specifically, we employed the p *x* y *x* m design (person *x* year *x* method) in which all survey respondents (p) provided their occupational title in all four survey years (y). Whether or not one was classified as an Engineer was determined by two methods (m), namely the best and self coded methods. The universe of admissible observations (UAO, Brennan, 1992a), therefore, contains 50,832 doctorate recipients who were ever employed in the Engineering profession between 1993 and 1999. For any particular survey year, an

Engineer received a value 1 if s/he was employed in Engineering and a 0 otherwise. The generalizability analysis allowed one to determine the extent to which (1) the professionals were employed the same number of years in Engineering; (2) the Engineering occupation employed a similar number of Ph.D.s across the survey years; (3) survey respondents reported their occupations as consistently as the objectively derived occupation; and (4) the interactions of these three factors.

Similar to Example 1, we estimated seven variance components (*p, y, m, py, pm, ym, pym,e*). Table 1 shows the estimates for the seven variance components and their corresponding standard errors. Both the weighted and unweighted methods yielded very similar results in the point estimate and the standard error of the variance components. For example, the ratio between the unweighted and weighted standard errors of the person effects was close to one because 0.00299 / 0.00296 = 1.0102 (i.e., $SE\left[\hat{\sigma}^2_p\right] \Big/ SE\left[\hat{\sigma}^{2\,(w)}_p\right]$).

Table 1: Comparisons of Variance Component Estimates (Weighted VS Unweighted)

	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_y$	$\hat{\sigma}^2_m$	$\hat{\sigma}^2_{py}$	$\hat{\sigma}^2_{pm}$	$\hat{\sigma}^2_{ym}$	$\hat{\sigma}^2_{pym,e}$
	person	*year*	*method*	*person by year*	*person by method*	*year by method*	*person by year by method, other errors*
Weighted	0.0675	0.0002	0.0008	0.0980	0.0047	0.0009	0.0477
Unweighted	0.0690	0.0002	0.0007	0.0969	0.0047	0.0008	0.0471
Ratio	1.0217	0.8984	0.9077	0.9888	0.9970	0.8485	0.9868
Weighted SE	0.0030	0.0006	0.0009	0.0021	0.0005	0.0006	0.0008
Unweighted SE	0.0030	0.0005	0.0008	0.0021	0.0005	0.0005	0.0008
Ratio	1.0102	0.8711	0.8940	0.9884	0.9893	0.8514	0.9868

Note: "Ratio" is the ratio of the unweighted estimates to the weighted estimates. The ratios were computed before the estimates were rounded to four decimal places.

Table 2 shows the percent contribution for each of the variance component estimates. The largest component was $\hat{\sigma}^2_{py}$ (0.098), which contributed to approximately 44.6% of the total variance in measuring occupational changes. Such results suggested that one can differentiate those who worked in the Engineering occupations for the same number of year by their job-switching patterns, where a job-switching pattern is characterized by the survey years in which a Ph.D. was employed in the Engineering profession as well as the years the doctorate was employed in other non-Engineering occupations (we summarize job switching patterns below and Chiu and Fecso, under review, offer an in-depth discussion). For example, two Ph.Ds. can be considered to have a different job-switching pattern even though they were both employed in an Engineering occupation for only one of the four survey years — hypothetically speaking, person A could work in an Engineering profession in 1993 but in a non-engineering profession in the subsequent years (the occupation pattern for person A would be [0 0 0 1], where the first, second, third, and fourth entries are binary variables for an Engineering employment in 1999, 1997, 1995, and 1993, respectively); person B could work in an non-engineering profession prior to becoming an

Engineer in 1999 (person B would have an occupation pattern [1 0 0 0]). Indeed, among the 487 doctorate recipients employed in Engineering for only one of the four survey years, 212 were employed in an Engineering occupation in only 1993; 90 were in only 1995; 61 were in only 1997; and 124 were in only 1999. The aforementioned differential job-switching pattern explained the relatively large $\hat{s}^2_{py}$.

Table 2: Comparisons of Variance Component Estimates Weighted VS Unweighted (Percent Contribution)

	$\hat{s}^2_p$	$\hat{s}^2_y$	$\hat{s}^2_m$	$\hat{s}^2_{py}$
Weighted	30.7%	0.1%	0.4%	44.6%
Unweighted	31.4%	0.1%	0.3%	44.2%

	$\hat{s}^2_{pm}$	$\hat{s}^2_{ym}$	$\hat{s}^2_{pyme}$
Weighted	2.1%	0.4%	21.7%
Unweighted	2.1%	0.4%	21.5%

The second large variance component estimate was $\hat{s}^2_p$, which indicated that, on average across all survey years and measurement methods, some Engineers had been employed in the profession for a longer duration than the others and the difference in duration accounted for approximately one third (30%) of the total job change variation.

Comparing the number of professionals employed in Engineering in different years can shed light in the stability of the occupation —— having a similar number of Engineers across different years can provide some evidence of stability whereas having a drastically different number of Engineers can provide some evidence of instability. The result that $\hat{s}^2_y$ accounted for only 0.1% of variation of the total job change suggested that the profession employed a similar number of Engineers in the survey years.

Like $\hat{s}^2_y$, the $\hat{s}^2_m$ accounted for only a small portion of total job change variation (0.4%) suggesting the objectively derived (best coding practice) and self-reported methods were relatively consistent in coding the Engineering profession. Resembling the $\hat{s}^2_y$ and the $\hat{s}^2_m$, the $\hat{s}^2_{ym}$ was relatively small suggesting that the two measurement methods were implemented consistently across the survey years.

The variance component estimate $\hat{s}^2_{pm}$, however, contributed to a larger share (2.1%) of the total variation than $\hat{s}^2_y$ and $\hat{s}^2_m$. One can interpret $\hat{s}^2_{pm}$ as an interaction between the variations due to person and method. It showed that the two occupational-determining methods were slightly more consistent for some survey respondents than the others but such differential consistency was relatively small comparing to the other sources of variation.

The person-by-year-by-method with any systematic and unsystematic variability $\hat{s}^2_{pyme}$ accounted for 21.7% of the total variation, suggesting that about one fifth of the job change variability in Engineering was due to: (a) the observation that Engineers changed jobs differentially in different survey years and the extent to which such a differential change occurred depends on which method was used to measure occupational titles; (b) any systematic variability such as the possibility that Engineers in some geographical regions were more mobile; and/ or (c) any unsystematic variability that was not measured.

Conclusion

The goal of incorporating sampling or survey weights into the framework of generalizability is to ensure that variance component are correctly estimated. The Chiu-Fecso method is designed for this purpose. In practice, the CFG method can be applied to educational assessment, psychological measurement, professional testing, and survey research where generalizability studies are called for to examine desirable variations and undesirable variations (measurement errors). Regardless of its dependence on sampling, the traditional G Theory framework assumes that simple random sampling is used. Indeed, national surveys and large-scale assessment programs use a variety of disproportional sampling techniques to ensure sample representations and account for non-responses. To this end a composite weight (final survey weight) is provided to analysts. Given that

the composite weight is frequently the only weighting information available to analysts, the current study extended the capacity of the G theory so that it can allow weights to be used.

In this article, we first introduced three principles in deriving the weighting method by showing how to estimate means and sums correctly. We then used the same principles to illustrate how to estimate variance components. Rules and step-by-step procedures were discussed. We validated the method using a published data set. The validation study suggested that weighted and unweighted variance component estimates can differ drastically if some cases receive a weight differ drastically from the others. Also, we showed that the traditional generalizability analysis is a special case of the weighted generalizability analysis. Two examples were provided to illustrate the applications of the weighting method in performance assessment and survey analysis. The weighted and unweighted variance component estimates of a large-scale operational data set yielded very similar conclusions.

Although the object of measurement, person, was the weighting facet in the two examples, this is not necessary to be the case. In practice, the weighting facet can be any facet in a crossed-two-faceted design (the main effect facets or the interaction effect facets). For instance, in standardized psychological or educational testing programs, researchers may desire to designate the item facet to be the weighting facet. This can be useful in examining the reliability of test scores when examinees do not respond to all items within the standard time. In the event that speededness happens, researchers can assign a lower weight to "not reached" items (those presented in the end of the test) than items presented in the beginning. Reese (1999) found that the true ability of low performing examinees is overestimated and that of high performing examinees is underestimated, when items are "locally dependent" or not reached by examinees (e.g., due to fatigue). The CFG method discussed in the current paper can be used to assign lower weights to not reached or locally dependent items. Future research can further investigate the extent to which different weights will change the reliability of test scores. Due to the page limits, it is not our intention to examine this topic in the current study.

Sometimes researchers are interested in assigning weights to multiple facets. For example, in educational assessment, one might be interested in oversampling minority students from the target population (i.e., weighting is used to adjust for the design effect). The weights to oversample minority students can be incorporated into a G study by assigning them to the facet related to persons (i.e., the object of measurement, Brennan, 1992a). In addition to assigning weights to the object of measurement, one can also weight the person-by-item facet. This can allow items to be weighted differently for individual students. Such an adaptive weighting mechanism can enable psychometricians to take into consideration the "opportunity to learn" when deciding the importance of an item on the test score. For example, one might assign a lower weight to an item when it is responded by a student whose school does not emphasize the learning objective of the item than when it is responded by another student who came from a school with a strong emphasis on the same item.

Similarly, in survey analysis, statisticians may desire to assign one set of weights to the sample of respondents and a completely different set of weights to the measurement methods. By doing so, survey statisticians could put a stronger emphasis on one measurement method (e.g., objective method) than the other (e.g., self-reported method) in evaluating quality of survey data. The aforementioned goal can be accomplished by developing a method to incorporate weighting schemes into multiple facets of a generalizability study (e.g., person and person-by-item). Future pursuit in developing a multifacet weighting scheme can apply the three principles discussed in the current study.

References

Adam, F., & Ujwal, K. (1999). Unmasking a phantom: A psychometric assessment of mystery shopping. *Journal of Retailing, 75*(2), 195-217.

Author. (2002). *Weighting Strategy*. [On-Line]: http://srsstats.sbe.nsf.gov/techinfo.html (Accessed Jan 25, 2002)

Bailar, B. A., Bailey, L., & Corby, C. (1978). A comparison of some adjustment and weighting procedures for survey data. In N. K. Namboodiri (Ed.), *Survey sampling and measurement* . New York: Academic Press.

Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation* (99-2). Princeton, NJ: Educational Testing Service.

Bell, J. F. (1985). Generalizability theory: The software problem. *Journal of Educational Statistics, 10* (1), 19-29.

Bennett, R., E. & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education, 9*, 133-150.

Biemer, P. P., & Fecso, R. S. (1995). Evaluating and controlling measurement error in business surveys. In B. Cox, Chinnappa, Christianson, Colledge, Kott. (Ed.), *Business survey methods* (257-281): Wiley & Sons, Inc.

Box, G. E.P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement, 27*, 93-108.

Brennan, R. L. (1992a). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Brennan, R. L. (1992b). NCME instructional module: Generalizability theory. *Educational measurement issues and practice, 11*(4), 27-34.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational measurement: issues and practice, 16*(4), 14-20.

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

Brennan, R. L, Gao, S., & Colton, D. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement, 55*(2), 157-176.

Brennan, R. L. (2000b). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24* (4), 339-353.

Brennan, R. L. (2001a). *Weights in generalizability theory*. Personal Communication in summer, 2001.

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

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement : Issues and Practice, 14* (4), 9-12,27.

Chiu, C. W. T., & Fecso, S. R. (in press review). *SEER: A graphical tool for multidimensional and categorical data*. *Journal of Data Science.*

Chiu, C. W. T., & Wolfe, E. W. (1997, April). *Generalizability theory: A new approach to analyze non-crossed performance assessment data.* Paper presented at the American Educational Research Association annual meeting, Chicago, IL.

Chiu, C. W. T., & Wolfe, E. W. (2002). A Method for Analyzing Sparse Data Matrices in the Generalizability Theory Framework**.** *Applied Psychological Measurement..26(3),* 319-336.

Chiu, C. W. T. (1999a). *Scoring performance assessments based on judgments: Utilizing meta-analysis to estimate variance components in generalizability theory for unbalanced situations.* Unpublished Dissertation. Michigan State University, Lansing, MI.

Chiu, C.W.T. (1999b, April). *Scoring performance assessments.* Poster for the Graduate Student Session at the 1999 Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

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

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24* (4), 310-324.

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics, 18* (1), 31-38.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York : Holt, Rinehart, and Winston.

Cronbach, L. J. (2000). *Weights in generalizability theory*. Personal Communication in summer, 2000.

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, NY: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generaliability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57* (3), 373-399.

Fyans, L. J. J. (Ed.). (1983). *Generalizability theory: Inferences and practical applications.* (Vol. 18): Jossey-Bass.

Hardy, L. P., & Eisenhower, D. L. (1994). *Developing methods for collecting and coding the occupation of persons with college degrees.* Paper presented at the Proceedings of the Section on Survey Research Methods. American Statistical Association, Alexandria, VA.

Holt, D. E., D. (1991). Methods of weighting for unit non-response. *The statistician, 40*, 333-342.

Johnson, S., & Bell, J. F. (1985). Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement, 22* (2), 107-119.

Kish (1995). *Survey sampling*. NY, New York: John Wiley & Son, Inc.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. London: Sage.

McGuinness, R. (1997). *1995 NSCG Coding Quality Evaluation.* . Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Millman, J., & Glass, G. V. (1967). Rules of thumb for writing the ANOVA table. *Journal of Educational Measurement, 4*(2), 41-51.

Rak, R., Chen, S., & Gray, L. (1997). *Occupation Coding: Best Coding and CATI Coding Methods.* (Research Compendium). Rockville, MD: Westat, Inc.

Rao, C. R. (1988). *Estimation of Variance Components and Applications*. New York: Elseuier Science.

Rao, J. N. K., & Sitter, R. R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In B. Lyberg, Collins, de Leeuw, Dippo, Schwarz, Trewin (Ed.), *Survey measurement and process quality* (pp. 753-768): John Wiley & Sons, Inc.

Rao, P. S. R. S. (1997). *Variance components estimation: Mixed models methodologies and applications*. New York: Chapman & Hall.

Reese, L. M. (1999). *Impact of local item dependence on item response theory scoring in CAT*. Computerized Testing Report 98-08. The Law School Admission Council. Newtown, PA.

Reiser, M., Fecso, R., & Chua, M. K. (1992). Some aspects of measurement error in the United States objective yield survey. *Journal of Official Statistics, 8* (3), 351-375.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82* (398), 387-394.

Sarndal, C. (1980). On pie-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika, 67* (3), 639-50.

Schott, J. R. (1997). *Matrix analysis for statistics*. New York: John Wiley & Sons, Inc.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York: Wiley.

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.

Shavelson, R. J., & Ruiz-Primo, M. A. (2000). On the psychometrics of assessing science understanding. In J. J. Mintzes & J. H. Wandersee (Eds.), *Assessing science understanding: A human*

constructivist view (303-341). San Diego: Academic Press, Inc.

Shipper, F. (1986). A study of four psychometric properties of the Jenkins Activity Survey Type. A scale with suggested modifications and validation. *Educational and Psychological Measurement, 46* (3), 551-64.

Suter, N., Harter, R., & Selfa, L. (1999). *1997 Survey of doctorate recipients methodology report.* Chicago, IL: National Opinion Research Center.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (41-53). New York: Russell Sage Foundation.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Appendix A: Derivations and Computational Examples for the Sum of Mean Scores used in Example One

Matrix notations were adopted from Scott (1997).

- **diag** is the operator to create a diagonal matrix.

- $\otimes$ is the Kronecker product operator, which multiplies the entire matrix in the right side of the operator to every element in the matrix to the left of the Kronecker operator. If **A** is an *m* x *n matrix* and **B** is a *p* x *q* matrix, then the Kronecker product of A and B, denoted $\mathbf{A} \otimes \mathbf{B}$, is the *mp* x *nq* matrix.

$$\begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & . & . & . & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & . & . & . & a_{2n}\mathbf{B} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & . & . & . & a_{mn}\mathbf{B} \end{bmatrix}$$

- We defined **w** and $\mathbf{w}_{f(p)}$ as row vectors. They are equivalent to the traditional matrix notation (Scott, 1997), which would define the two row vectors as transposes (i.e., $\mathbf{w}^T$ and $\mathbf{w}^T_{f(p)}$).

- $\odot$, the Hadamard operator, is the elementwise multiplication operator for two matrices. The traditional Hadamard operator $\odot$ requires that two quantities to be expressed separately in the left and in the right sides of the operator. This becomes cumbersome when the two quantities are identical, because one would have to repeat a quantity twice. For example, to perform an elementwise multiplication of $(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \underset{n_r \text{ x } n_r}{\mathbf{I}} \otimes \underset{n_i \text{ x } 1}{\mathbf{1}} \bullet (1/n_i))$ to itself, one would write: $(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \underset{n_r \text{ x } n_r}{\mathbf{I}} \otimes \underset{n_i \text{ x } 1}{\mathbf{1}} \bullet (1/n_i)) \odot (\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \underset{n_r \text{ x } n_r}{\mathbf{I}} \otimes \underset{n_i \text{ x } 1}{\mathbf{1}} \bullet (1/n_i))$. To save space, we defined a parsimonious version of the Hadamard operator, to represent an elementwise power multiplication. For example, $\mathbf{X}^{\odot 2}$ indicates that the elements in **X** were raised to the second

power. Using the new operator, the aforementioned cumbersome notation can be simplified as follows. $(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \underset{n_r \text{ x } n_r}{\mathbf{I}} \otimes \underset{n_i \text{ x } 1}{\mathbf{1}} \bullet (1/n_i))^{\odot 2}$. In summary, $\mathbf{X}^{\odot 2} = \mathbf{X} \odot \mathbf{X}$.

- In each of the following equations, the first line shows the summation notation of the sums of squared means and the second line shows the matrix notation of the same quantity.

$$\sum_p \bar{x}^2_{p..} = \sum_p w_p \left(\left(\frac{1}{n_i n_r} \sum_i \sum_r x_{pir}\right)^2\right) = \mathbf{w} \bullet \text{diag}\left(\left(\mathbf{X} \bullet \underset{n_{ir} \text{ x } 1}{\mathbf{1}}\right) \bullet (1/n_{ir})\right) \bullet \left(\left(\mathbf{X} \bullet \underset{n_{ir} \text{ x } 1}{\mathbf{1}}\right) \bullet (1/n_{ir})\right) \quad (13)$$

$$\sum_i \bar{x}^2_{i..} = \sum_i \left(\sum_p w_{f(p)} \left(\frac{1}{n_r} \sum_r x_{pir}\right)\right)^2 = \left(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \begin{bmatrix} \underset{n_i \text{ x } n_i}{\mathbf{I}} \\ \underset{n_i \text{ x } n_i}{\mathbf{I}} \end{bmatrix} \bullet (1/n_r) \bullet \text{diag}\left(w_{f(p)} \bullet \mathbf{X} \bullet \begin{bmatrix} \underset{n_i \text{ x } n_i}{\mathbf{I}} \\ \underset{n_i \text{ x } n_i}{\mathbf{I}} \end{bmatrix} \bullet (1/n_r)\right)\right) \bullet \underset{n_i \text{ x } 1}{\mathbf{1}} \quad (14)$$

$$\sum_r \bar{x}^2_{r..} = \sum_r \left(\sum_p w_{f(p)} \left(\frac{1}{n_i} \sum_i x_{pir}\right)\right)^2 = \left(\mathbf{w}_{f(p)} \bullet \mathbf{X} \bullet \underset{n_r \text{ x } n_r}{\mathbf{I}} \otimes \underset{n_i \text{ x } 1}{\mathbf{1}} \bullet (1/n_i)\right)^{\odot 2} \bullet \underset{n_r \text{ x } 1}{\mathbf{1}} \quad (15)$$

$$\sum_p \sum_i \bar{x}^2_{pi.} = \sum_p w_p \left(\sum_i \left(\frac{1}{n_r} \sum_r x_{pir}\right)^2\right) = \mathbf{w} \bullet \left(\left(\mathbf{X} \bullet \left(\underset{n_r \text{ x } 1}{\mathbf{1}} \otimes \underset{n_i \text{ x } n_i}{\mathbf{I}}\right)\right) \bullet (1/n_r)\right)^{\odot 2} \bullet \underset{n_r \text{ x } 1}{\mathbf{1}} \quad (16)$$

$$\sum_p \sum_r \bar{x}^2_{pr.} = \sum_p w_p \left(\sum_r \left(\frac{1}{n_i} \sum_i x_{pir}\right)^2\right) = \mathbf{w} \bullet \left(\mathbf{X} \bullet \left(\underset{n_r \text{ x } n_r}{\mathbf{I}} \otimes \underset{n_i \text{ x } 1}{\mathbf{1}}\right) \bullet (1/n_i)\right)^{\odot 2} \bullet \underset{n_r \text{ x } 1}{\mathbf{1}} \quad (17)$$

$$\sum_i \sum_r \bar{x}^2_{ir.} = \sum_i \sum_r \sum_p \left(w_{f(p)} x_{pir}\right)^2 = \left(\mathbf{w}_{f(p)} \bullet \mathbf{X}\right)^{\odot 2} \bullet \underset{n_{ir} \text{ x } 1}{\mathbf{1}} \quad (18)$$

$$\sum_p \sum_i \sum_r x^2_{pir} = \sum_p w_p \sum_i \sum_r x^2_{pir} = \mathbf{w} \bullet \mathbf{X}^{\odot 2} \bullet \underset{n_{ir} \text{ x } 1}{\mathbf{1}} \quad (19)$$

$$\bar{x}^2 = \left(\frac{1}{n_p n_i n_r} \sum_p w_p \sum_i \sum_r x_{pir}\right)^2 = \left((1/n_{pir}) \bullet \mathbf{w} \bullet \mathbf{X} \bullet \underset{n_{ir} \text{ x } 1}{\mathbf{1}}\right)^{\odot 2} \quad (20)$$

$$\underset{n_p \text{x } n_{ir}}{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

$$\underset{1 \text{x } n_p}{\mathbf{w}_{\text{p}}} = \begin{bmatrix} 0.8 & 1.2 & 1.6 & 0.4 \end{bmatrix} \tag{22}$$

$$\underset{1 \mathbf{x} n_p}{\mathbf{w}_{\text{f(p)}}} = \begin{bmatrix} 0.2 & 0.3 & 0.4 & 0.1 \end{bmatrix} \tag{23}$$

$$1/n_i = 1/3 \tag{24}$$

$$1/n_r = 1/2 \tag{25}$$

$$1/n_{ir} = 1/6 \tag{26}$$

$$1/n_{pir} = 1/24 \tag{27}$$

$$\underset{n_i \text{ x} 1}{1} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^{\text{T}} \tag{28}$$

$$\underset{n_r \text{ x} 1}{1} = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\text{T}} \tag{29}$$

$$\underset{n_{ir} \text{ x } 1}{1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^{\text{T}} \tag{30}$$

$$\underset{n_i \text{ x } n_i}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{31}$$

$$\underset{n_r \text{ x } n_r}{\mathbf{I}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{32}$$

$$\mathbf{X}^{\odot 2} = \begin{bmatrix} 1^2 & 1^2 & 1^2 & 0^2 & 0^2 & 0^2 \\ 0^2 & 1^2 & 1^2 & 1^2 & 0^2 & 1^2 \\ 1^2 & 0^2 & 0^2 & 0^2 & 1^2 & 1^2 \\ 0^2 & 0^2 & 0^2 & 0^2 & 0^2 & 1^2 \end{bmatrix} = \text{X} \odot \text{X} \tag{33}$$

$$\underset{n_r \times n_r}{\mathbf{I}} \otimes \underset{n_i \times 1}{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad (34)$$

By substituting (21) through (34) into the corresponding elements in (13) through (20), following results are obtained and used to compute the weighted sum of squared mean shown in (10).

$$\sum_p \bar{x}^2_{p..} = 1.444 \quad (35)$$

$$\sum_i \bar{x}^2_{i..} = 0.8275 \quad (36)$$

$$\sum_r \bar{x}^2_{r..} = 0.5344 \quad (37)$$

$$\sum_p \sum_i \bar{x}^2_{pi.} = 3.7000 \quad (38)$$

$$\sum_p \sum_r \bar{x}^2_{pr.} = 2.8000 \quad (39)$$

$$\sum_i \sum_r \bar{x}^2_{ir.} = 1.7500 \quad (40)$$

$$\sum_p \sum_i \sum_r x^2_{pir} = 12.400 \quad (41)$$

$$\bar{x}^2 = 0.2669 \quad (42)$$

Appendix B: A Comparison between the Traditional Unweighted and the Chiu-Fecso Weighted Methods.

	Unweighted VC Brennan, 1992 (p.38)	Unweighted VC Chiu-Fecso Unit Weights	Weighted VC Chiu-Fecso Disproportionate Weights	Ratio: Weighted VC / Unweighted VC
P	0.5528	0.5528	0.9634	1.7428
I	0.4417	0.4417	0.5656	1.2805
R	0.0074	0.0074	0.0221	2.9865
Pi	0.575	0.575	0.4432	0.7708
Pr	0.1009	0.1009	0.0349	0.3459
Ir	0.1565	0.1565	0.0562	0.3591
pir,e	0.9352	0.9352	0.5776	0.6176

Notes: Unit weights: w = [1 1 1 1 1 1 1 1 1 1] , Disproportionate weights: w = [1 1 1 1 1 1 1 1 1 10]. Data source: Brennan (1992a, p.38).

Appendix C: Weighted Variance Component Estimates by Age Group for Example Two.

Variance Component Estimates

Age Group	Below 30	35-39	40-44	45-49	50-54	55-59	60-64	65-69	Above 70
p	0.1667	0.055	0.0658	0.0753	0.0732	0.0661	0.0677	0.0561	0.0398
y	0	0.001	0.0002	0	0	0	0	0.0157	0.0203
m	0	0.0003	0.0003	0.0007	0.0008	0.0013	0.0015	0.0002	0.0002
py	0.0833	0.0848	0.0805	0.0943	0.0865	0.0973	0.0972	0.1314	0.1528
pm	0	0.0022	0.0029	0.0029	0.0064	0.0076	0.0067	0.001	0.0056
ym	0	0.0002	0.0001	0.0006	0.0011	0.0013	0.0015	0.001	0.0025
pym,e	0	0.0332	0.044	0.0509	0.0525	0.0523	0.0475	0.0425	0.0348

p: person, *y*: year, *m*: method; *py* = person by year; *pm*: person by method, *ym*: year by method, *pym,e*: person by year by method and other errors.

Percent Contribution

Age Group	Below 30	35-39	40-44	45-49	50-54	55-59	60-64	65-69	Above 70
p	66.7%	31.1%	34.0%	33.5%	33.2%	29.3%	30.5%	22.6%	15.5%
y	0.0%	0.6%	0.1%	0.0%	0.0%	0.0%	0.0%	6.3%	7.9%
m	0.0%	0.1%	0.2%	0.3%	0.4%	0.6%	0.7%	0.1%	0.1%
py	33.3%	48.0%	41.5%	42.0%	39.2%	43.1%	43.8%	53.0%	59.7%
pm	0.0%	1.3%	1.5%	1.3%	2.9%	3.4%	3.0%	0.4%	2.2%
ym	0.0%	0.1%	0.0%	0.3%	0.5%	0.6%	0.7%	0.4%	1.0%
pym,e	0.0%	18.8%	22.7%	22.6%	23.8%	23.2%	21.4%	17.1%	13.6%
Sample Size	3	202	439	400	440	392	256	140	116