Generalizability Theory With One-Facet Nonadditive Models

Applied Psychological Measurement
2016, Vol. 40(6) 367–386
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616651603
apm.sagepub.com

Jinming Zhang¹ and Chih-Kai (Cary) Lin²

Abstract

In generalizability theory (G theory), one-facet models are specified to be additive, which is equivalent to the assumption that subject-by-facet interaction effects are absent. In this article, the authors first derive estimators of variance components (VCs) for nonadditive models and show that, in some cases, they are different from their counterparts in additive models. The authors then demonstrate and later confirm with a simulation study that when the subject-by-facet interaction exists, but the additive-model formulas are used, the VC of subjects is underestimated. Consequently, generalizability coefficients are also underestimated. Thus, depending on the nature of interaction effects, an appropriate model, either additive or nonadditive, should be used in applications of G theory. The nonadditive G theory developed in this article generalizes current G theory and uses data at hand to determine when additive or nonadditive models should be used to estimate VCs. Finally, the implications of the findings are discussed in light of an analysis of real data.

Keywords

additivity index, generalizability theory, negative estimated variance component, repeated measures, within-subject design, Tukey's test, nonadditivity

Introduction

Generalizability theory (G theory) investigates the dependability or reliability of the results from a sample of behavioral measurements at hand, and the results are then to be generalized to any valid measurements for decision making (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Nageswari, & Gleser, 1963; Shavelson & Webb, 1991). It aims to pinpoint all potential sources (specified by users) of variation in measurement scores by using ANOVA. The variance components (VCs) of identified sources of variation are the basic building blocks of G theory. An analysis based on G theory typically consists of two steps, a generalizability study and a decision study. A decision can be a relative or an absolute decision. Generalizability coefficients (G coefficients) measure the dependability or reliability of results

¹University of Illinois at Urbana–Champaign, IL, USA

²Center for Applied Linguistics, Washington, DC, USA

Corresponding Author:

Jinming Zhang, Department of Educational Psychology, University of Illinois at Urbana–Champaign, 1310 South Sixth Street, 236A Education Building, Champaign, IL 61820, USA.

Email: jmzhang@illinois.edu

that can be generalized from a sample of observations to a universe of admissible observations. For details about G theory, see Brennan (2001) and Shavelson and Webb (1991).

Model specifications in current G theory assume additivity such that “all effects in the model are uncorrelated” (Brennan, 2001, p. 23), which is statistically equivalent to assuming no interaction between the subject (e.g., person or essay) and the facet (e.g., item or rater) in a design with one facet. However, the assumption of additivity may not be satisfied in practice. In a multiple-rating design of performance-based assessment, each subject response is scored by multiple raters. Clearly, the subject-by-facet interaction may be present. For example, such an interaction may be introduced when some raters tend to award higher scores for certain responses, whereas other raters award lower scores for such responses. Another scenario that may also introduce interaction is when some raters tend to give polarized scores (very high scores for high performance students and very low scores for low performance students) in a performance-based assessment, whereas other raters tend to give toward-average scores.

In the presence of interaction, directly applying additive models in the analyses and failing to consider the interaction effects may lead to underestimation of the VC of subjects as the article will reveal later. Depending on the degree to which the VC is underestimated, negative estimated VCs may occur. Consequently, generalizability coefficients are also underestimated and may even be negative as will become clear in the next section.

Although the assumption of additivity has been widely accepted in G-theory applications, a more general linear model is the nonadditive model, which relaxes the assumption of no interaction (i.e., additivity). In this article, the differences and implications for additive and nonadditive model specifications are discussed. Nonadditive models are introduced to G theory to broaden the scope of current G-theory applications such that nonadditivity is also considered. The authors further show that the estimated VCs of subjects are different for the additive and nonadditive models. Thus, appropriate uses of the formulas rest on careful examinations of the underlying model assumptions and of the nature of subject-by-facet interaction effects.

The rest of the article is organized as follows: First, the difference between random and fixed effects is discussed, and the nonadditive model is introduced. Next, the estimators of VCs for one-facet nonadditive models, which include an additive model as a special case, are derived. Tukey’s test is introduced for testing nonadditivity, and a nonadditive G-theory approach is developed. Then, two simulation studies are conducted. The first simulation study shows that the nonadditive approach produces almost the same accurate results as the additive approach used in current G theory when data contain no interaction (i.e., an additive model is true). The second simulation study confirms that when the subject-by-facet interaction is present, the formulas for estimated VCs following the additive models can severely underestimate the VC for subjects, especially when the number of facet levels is small, but the results from the nonadditive approach are much improved. A real data set with one negative VC estimate based on current additive G theory is reanalyzed using the formulas for estimated VCs in nonadditive G theory developed in this article. The results show that all VC estimates are now positive as a result of using the nonadditive approach. Finally, the nonadditive approach and its implications on the G-theory framework are further discussed.

Models and VC Estimates

Suppose in a writing-proficiency study, each of 25 students writes an essay on the same topic, and each essay is graded by all three raters. This is a fully crossed $p \times i$ design, where p refers to the subject (e.g., person or essay) and i refers to a facet (e.g., item or rater) in G theory. Clearly, given that a single essay is graded by multiple raters, this one-facet design is a *repeated measure* design or a *within-subject* design (see Crocker & Algina, 2008; Everitt, 1995).

The subject effects are considered to be random in G theory, while the facet can be either fixed or random. Thus, the linear model used in G theory is either a random-effect or mixed-effect model. The difference between random and fixed facets is that the former treats, for example, the rater facet as random and therefore the generalizability is to a universe of similarly qualified raters, whereas the latter treats the rater facet as fixed and hence the derived inferences are limited to the raters recruited in measurement design at hand. For details about guidelines for judging whether random or fixed effects should be used in a general linear model, see Gelman (2005, pp. 21-23), Kreft and de Leeuw (1998, Section 1.3.3), and Snijder and Bosker (2012, Section 4.3.1).

A random-effect linear model for a one-facet design in G theory is

$$\mathbf{X}_{pi} = \mu + \tau_p + \alpha_i + (\tau\alpha)_{pi} + \varepsilon_{pi}, \quad (1)$$

where μ is the grand mean in the population of subjects and universe of admissible observations contains the facet, and τ_p and α_i are the subject and facet effects, respectively. The two effects are assumed to be uncorrelated and identically distributed with mean zeros and variances σ_p^2 and σ_i^2 , respectively. Here the same letter is used to represent a factor and an index of a factor level. In Equation 1, ε_{pi} is a random error component specific to subject p at facet level i . These error terms $\{\varepsilon_{pi}\}$ are assumed to be uncorrelated and identically distributed with mean zero and variance σ_e^2 , and uncorrelated with all other terms on the right side of Equation 1. Under these assumptions, Scheffe (1959) proved that $\{(\tau\alpha)_{pi}\}$ are completely uncorrelated among themselves, and uncorrelated with other components on the right side of Equation 1. This result is summarized as the following theorem:

Theorem 1: In the Random-Effect Linear Model 1, the effects $\{\tau_p\}$, $\{\alpha_i\}$, and $\{(\tau\alpha)_{pi}\}$ are completely uncorrelated. Under the normality assumption that all effects $\{\tau_p\}$, $\{\alpha_i\}$, $\{(\tau\alpha)_{pi}\}$, and $\{\varepsilon_{pi}\}$ are jointly normal, these effects are completely independent.

Note that Scheffe (1959) considered all complete two-way layouts. This includes the special case where there is only one observation per cell, which is the case considered in G theory and in this article.

According to Scheffe's theorem, the interaction term $(\tau\alpha)_{pi}$ has exactly the same characteristics as the error term ε_{pi} . Therefore, $\{\varepsilon_{pi}^* = (\tau\alpha)_{pi} + \varepsilon_{pi}\}$ are uncorrelated and identically distributed with mean zero and variance $\sigma_{e^*}^2$, and are uncorrelated with the other terms on the right side of Equation 1, where $\sigma_{e^*}^2 = \sigma_{pi}^2 + \sigma_e^2$. That is, these two components, $(\tau\alpha)_{pi}$ and ε_{pi} , can be combined together and regarded as a single pure error term. In other words, the interaction term presented in the model is superfluous. There is no (genuine) interaction effect in the random-effect linear model and the interaction component can be removed from the random-effect linear model. The random-effect linear model (1) is *additive*. Thus, the following result is obtained:

Corollary 1: The Random-Effect Linear Model 1 is an additive model or has no genuine interaction.

It is a common belief that the subject-by-facet interaction and the pure error are totally indistinguishable because there is only one observation per cell in a one-facet design. Thus, these two components are always assumed to be uncorrelated with other components and combined together as a single error term in Model 1 in current G theory. That is, current G theory only uses random-effect additive models in one-facet designs and assumes there is no subject-by-facet interaction.

Tabachnick and Fidell (2007) pointed out that

[o]ne's fondest hope in repeated measures is that there is no genuine (population) interaction, so that the error term is only random error. This is a forlorn hope with many real data sets, because of contingencies in the responses of the same cases as they are measured repeatedly over time. (p. 248)

The (genuine) interaction effects, if exist, are correlated among themselves and/or correlated with other main effects and cannot be treated as error components in statistics. One type of interaction is that $(\tau\alpha)_{pi} = \eta\tau_p\alpha_i$, where η is a constant scale factor. In practice, genuine interaction effects may exist and are typically associated with particular levels of other effects. In a writing-proficiency study, for example, if some raters tend to give polarized scores, essay-by-rater interaction effects exist, and such effects belong to idiosyncratic ratings given by specific raters to particular essays. When the interaction terms $\{(\tau\alpha)_{pi}\}$ are correlated among themselves and/or correlated with some other terms on the right side of Equation 1, Model 1 is called *nonadditive*.

In current G theory, whether a facet is treated as random or fixed is determined by the objective of a research study and thus, a facet is assumed to be random when a user intends to generalize results toward the underlying universe. In a writing-proficiency study, although raters are typically not randomly selected from the universe of all qualified raters, they are generally considered interchangeable with other raters in the universe. Thus, the rater facet is regarded as random (see Shavelson & Webb, 1991). In addition, the facet is always considered random in a one-facet design in current applications of G theory. However, this practice is not appropriate in certain scenarios (i.e., in the presence of interaction). The issue occurs due to the potential incongruence between actually selected facet levels (conditions of measurement, for example, raters) and the universe of admissible observations (e.g., all qualified raters). For example, to treat the rater facet as random in a writing-proficiency study, one crucial requirement for qualified raters is that they are able to grade students' essays objectively (no interaction) according to Corollary 1. In reality, however, "qualified raters" are typically identified according to their academic degrees and work experience, and these qualifications do not necessarily guarantee objectivity in scoring. That is, students' scores may be interfered with by such a nominally qualified rater's preferences and thus, subject-by-facet interaction may exist. In such a case, it is not appropriate to use the random-effect additive model due to the interaction. In short, although the facet is conceptualized as random (no interaction) in a one-facet design, whether the facet can be treated as such in a statistical model is an empirical question, and therefore one cannot take it for granted in practice. From a statistical perspective, users need to check whether interaction effects exist in the data at hand, especially in a design where the subject and/or facet involve humans, before they can safely treat the facet as random in a one-facet linear model.

Three research issues should be investigated. First, given a data set from a one-facet design, is there any method to test whether subject-by-facet interaction effects are significant or not? Clearly, if the interaction effects are not significant, one can treat the facet as random and use current G theory to complete the rest of the data analyses. However, if the interaction effects are significant, then the next research question is "What are the consequences of continuing to treat the facet as random and applying current G theory to the data?" Finally, if the consequences are severe, one needs to develop an appropriate new method to deal with possibly existing interaction effects under a more general setting than current G theory.

According to Corollary 1, a random-effect linear model is additive and assumes no interaction in a one-facet design. As such, for a model to accommodate interaction effects, the facet has to be assumed fixed and thereby a mixed-effect nonadditive model should be applied.

Model 1 becomes a mixed-effect linear model when the facet is treated as fixed: The treatment effects α_i are unknown parameters, $\sum \alpha_i = 0$, and the pseudo variance is defined as

$$\sigma_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} \alpha_i^2,$$

where n_i is the number of levels of the facet. For convenience purposes, the same notation is used in this article for the pseudo variance as the facet variance in a random-effect linear model. The interaction component $(\tau\alpha)_{pi}$ is still random with a constraint that $\sum_i (\tau\alpha)_{pi} = 0$. It must be emphasized here that a mixed-effect linear model can be either additive or nonadditive, while a random-effect linear model must be additive. In a nonadditive model, $(\tau\alpha)_{pi}$ and $(\tau\alpha)_{pi'}$ and τ_p and $(\tau\alpha)_{pi}$ may be correlated, whereas they are uncorrelated (and independent under normality assumption) in an additive model. When a mixed-effect linear model is additive in the sense that $(\tau\alpha)_{pi}$ and $(\tau\alpha)_{pi'}$, and τ_p and $(\tau\alpha)_{pi}$ are uncorrelated with each other, the interaction term $(\tau\alpha)_{pi}$ in Equation 1 is *removable* and can be absorbed completely in the corresponding error term ε_{pi} as shown before for a random-effect model. Thus, an additive model with a fixed or random facet can be rewritten as

$$X_{pi} = \mu + \tau_p + \alpha_i + \varepsilon_{pi}. \quad (2)$$

That is, the Additive Model 2 is equivalent to Model 1 when the facet is random or when the facet is fixed without any interaction. The advantage of using Equation 2 for an additive case is that it explicitly shows that no genuine interaction is considered in the model, whereas Model 1 may give readers the wrong impression that the interaction is always considered.

Corollary 2: (a) The Mixed-Effect Linear Model 1 can be either additive or nonadditive (either without or with genuine interaction). (b) To consider genuine interaction effects, the facet should be treated as fixed and a mixed-effect linear model should be used.

The regular notations of ANOVA are used in this article: The sum squares (SS) are

$$SS_p = \sum_{pi} (\bar{X}_{p.} - \bar{X}_{..})^2 = n_i \sum_p (\bar{X}_{p.} - \bar{X}_{..})^2,$$

$$SS_i = \sum_{pi} (\bar{X}_{.i} - \bar{X}_{..})^2 = n_p \sum_i (\bar{X}_{.i} - \bar{X}_{..})^2,$$

$$SS_{pi,e} = \sum_{pi} (X_{pi} - \bar{X}_{p.} - \bar{X}_{.i} + \bar{X}_{..})^2.$$

The mean squares (MS) are obtained by dividing these SS by their corresponding degrees of freedom (df).

The expected MS for a one-facet model under the additive (for both random and mixed) and nonadditive (mixed) models are presented in Table 1. As the same notation is used, the formulas for both random- and mixed-effect models are the same when models are additive. Given a set of data, all estimation results based on an additive model with a fixed facet are the same as those with a random facet in any one-facet design. The only difference is in the interpretation of the analysis results. For readers' convenience, Table 1 includes expected MS for the additive models under both Equations 1 and 2. Readers can easily find that the column for the Additive Model 1 in Table 1 is exactly the same as Table 3.1 in Shavelson and Webb (1991). The

Table 1. Expected MS in a One-Facet Design.

Sources of variation	MS	Expected MS		
		Additive Model 1 with a fixed or random facet	Additive Model 2 with a fixed or random facet	Nonadditive Model 1 with a fixed facet
Person (<i>p</i>)	MS_p	$\sigma_e^2 + \sigma_{pi}^2 + n_i \sigma_p^2$	$\sigma_e^2 + n_i \sigma_p^2$	$\sigma_e^2 + n_i \sigma_p^2$
Item (<i>i</i>)	MS_i	$\sigma_e^2 + \sigma_{pi}^2 + n_p \sigma_i^2$	$\sigma_e^2 + n_p \sigma_i^2$	$\sigma_e^2 + \sigma_{pi}^2 + n_p \sigma_i^2$
<i>pi</i> , <i>e</i>	$MS_{pi,e}$	$\sigma_e^2 + \sigma_{pi}^2$	σ_e^2	$\sigma_e^2 + \sigma_{pi}^2$

Note. MS = mean squares.

formulas for the expected MS for a within-subject design can also be found in experimental design books (e.g., Davis, 2002; Keppel & Wickens, 2004; Myers, 1979; Myers, Well, & Lorch, 2010; Scheffe, 1959).

The formulas for estimated VCs under the Nonadditive Mixed-Effect Model 1 can be obtained by replacing the expected MS with the corresponding MS in Table 1 (i.e., by the moment matching method). Specifically, according to the last column of Table 1, let

$$MS_p = \hat{\sigma}_e^2 + n_i \hat{\sigma}_p^2,$$

$$MS_i = \hat{\sigma}_{pi,e}^2 + n_p \hat{\sigma}_i^2 = \hat{\sigma}_e^2 + \hat{\sigma}_{pi}^2 + n_p \hat{\sigma}_i^2,$$

$$MS_{pi,e} = \hat{\sigma}_{pi,e}^2 = \hat{\sigma}_e^2 + \hat{\sigma}_{pi}^2.$$

By solving these three equations, the formulas for estimated VCs under the Nonadditive Mixed-Effect Model 1 are obtained:

$$\hat{\sigma}_{pi,e}^2 = MS_{pi,e}. \quad (3)$$

$$\hat{\sigma}_i^2 = \frac{MS_i - MS_{pi,e}}{n_p}. \quad (4)$$

$$\hat{\sigma}_p^2 = \frac{MS_p - MS_{pi,e} + \hat{\sigma}_{pi}^2}{n_i} = \frac{MS_p - \hat{\sigma}_e^2}{n_i}. \quad (5)$$

In the case of an additive model, according to the fourth column of Table 1, σ_p^2 can be estimated by

$$\hat{\sigma}_p^2 = \frac{MS_p - MS_{pi,e}}{n_i}, \quad (6)$$

and the estimators of $\sigma_{pi,e}^2$ and σ_i^2 are the same as Equations 3 and 4. One can observe that these three estimated VCs for an additive random- or mixed-effect model (Equations 3, 4, and 6) are exactly the same as those currently used in G theory (see Equations 3.7, 3.8, and 3.9 of Shavelson & Webb, 1991). When the model is additive (i.e., no interaction), Equations 5 and 6 are exactly the same.

Corollary 3: (a) The three estimated VCs are the same between the additive model with a fixed facet and the additive model with a random facet in a one-facet design. (b) In the case

of a nonadditive model (i.e., mixed-effect model with a fixed facet), the estimated VC for subjects (Equation 5) is different from the corresponding one (Equation 6) for the additive models, whereas the other two estimated VCs are the same for both the additive and nonadditive models.

The variances of interaction and pure error are confounded with each other in a design with only one observation in each cell. In general, σ_e^2 and σ_{pi}^2 cannot be estimated separately (Cronbach et al., 1972). Consequently, $\hat{\sigma}_p^2$ in Equation 5 is also not readily calculated in a nonadditive model. Thus, the key issue for the use of nonadditive models is how to estimate σ_e^2 and σ_{pi}^2 separately.

Tukey (1949) proposed a single-degree-of-freedom test for nonadditivity (interaction) in the two-way layout (i.e., one-facet design) with only one observation per cell. Note that the number of df for the residual sum of squares, $SS_{pi,e}$, is $df_{pi,e} = (n_p - 1)(n_i - 1)$. Tukey partitioned $SS_{pi,e}$ into two components: a one-degree-of-freedom interaction component due to nonadditivity, SS_{nonadd} , and the remaining component, SS_{remain} , as an error term for hypothesis testing of nonadditivity. Specifically,

$$SS_{nonadd} = \frac{\left[\sum_{pi} X_{pi} (\bar{X}_{p.} - \bar{X}_{..}) (\bar{X}_{.i} - \bar{X}_{..}) \right]^2}{\sum_p (\bar{X}_{p.} - \bar{X}_{..})^2 \sum_i (\bar{X}_{.i} - \bar{X}_{..})^2}$$

and $SS_{remain} = SS_{pi,e} - SS_{nonadd}$. Thus, $MS_{nonadd} = SS_{nonadd}$ can be regarded as an estimate of σ_{pi}^2 , and $MS_{remain} = SS_{remain}/(df_{pi,e} - 1) = SS_{remain}/(n_p n_i - n_p - n_i)$ as an estimate of σ_e^2 . As such, Tukey's method can be used to estimate σ_e^2 and σ_{pi}^2 separately. The F ratio statistic of Tukey's test for nonadditivity is $F = MS_{nonadd}/MS_{remain}$, with 1 and $n_p n_i - n_p - n_i$ degrees of freedom (see Kirk, 2013; Myers, 1979).

Myers (1979) showed that Tukey's test is remarkably sensitive to average-by-effect interactions, but may not work in cases where some interaction effects were clearly present. Kirk (2013) also wrote that

Tukey's test is sensitive to situations in which the scores for each block follow the same general trend, but the amount of change from a_1 to a_2 , a_2 to a_3 , and so on is not the same for all blocks. This is a limitation of the test because other forms of nonadditivity do occur. Nevertheless, it can be a useful preliminary test in designs that have one score per cell. (p. 300)

Myers (1979) presented a plot to summarize the pattern of interaction to which Tukey's test is sensitive, and pointed out that "the nature of most interaction effects is such that the cross-product row-mean relation will have a linear component and Tukey's test will therefore be serviceable" (see Myers, 1979, p. 186, Figures 7-2). Although Tukey's test has different powers for different patterns of interaction, it can control the rate of Type I errors at any significance level, which lends credence to using Tukey's test in G-theory settings (see Lin, 2014).

Let

$$\lambda^2 = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_{pi}^2} = \frac{\sigma_e^2}{\sigma_{pi,e}^2} = 1 - \frac{\sigma_{pi}^2}{\sigma_{pi,e}^2},$$

where $0 \leq \lambda^2 \leq 1$, which is called the *additivity index* in this article. Note that $\lambda^2 = 1$ if, and only if, the model is additive; that is, the index measures the degree to which nonadditivity exists in data: the smaller the value is, the stronger the nonadditivity is present.

The Nonadditive Approach

In this article, a conservative approach is adopted: First, Tukey's test for nonadditivity is used to gauge whether any interaction exists: If $F = (MS_{\text{nonadd}}/MS_{\text{remain}}) \leq F_{.05}(1, n_p n_i - n_p - n_i)$ (i.e., Tukey's test is not significant at the significance level 0.05), retain the null hypothesis of additivity (i.e., $\lambda^2 = 1$ and an additive model will be used); otherwise, a nonadditive model will be used and λ^2 can be estimated as either

$$\hat{\lambda}^2 = \frac{MS_{\text{remain}}}{MS_{pi,e}}$$

or

$$\hat{\lambda}^2 = 1 - \frac{MS_{\text{nonadd}}}{MS_{pi,e}}.$$

The above two estimates are not necessarily the same. In this article, the average of these two estimates is used as the default value for estimated λ^2 .

It should be noted that this nonadditive approach uses a nonadditive model only when Tukey's test is significant. Thus, when applied to data without any interaction, the nonadditive approach is the same as the additive approach used in current G theory except for the cases where Tukey's test makes Type I errors. When interaction effects are not significant, the facet can be treated as random.

Clearly, $\hat{\sigma}_e^2 = \hat{\lambda}^2 \hat{\sigma}_{pi,e}^2$ and $\hat{\sigma}_{pi}^2 = (1 - \hat{\lambda}^2) \hat{\sigma}_{pi,e}^2$. That is, one can estimate σ_e^2 and σ_{pi}^2 separately in the design with only one observation per cell by using Tukey's method. From Equation 5, the estimated VC for subjects becomes

$$\hat{\sigma}_p^2(\hat{\lambda}^2) = \frac{MS_p - \hat{\lambda}^2 \hat{\sigma}_{pi,e}^2}{n_i} = \frac{MS_p - \hat{\lambda}^2 MS_{pi,e}}{n_i}. \quad (7)$$

By comparing Equation 6 of an additive model with Equation 7 of a nonadditive model, one can see that G theory currently only uses additive models in which $\lambda^2 = 1$. That is, $\hat{\sigma}_p^2 = \hat{\sigma}_p^2(\text{additive}) = \hat{\sigma}_p^2(\hat{\lambda}^2 = 1)$ is a special estimator of σ_p^2 based on an additive model. Note that the primary difference between Equations 6 and 7 is that the former assumes no subject-by-facet interaction (i.e., $\lambda^2 = 1$), whereas the latter explicitly models such interactions ($0 \leq \lambda^2 \leq 1$) with no interaction ($\lambda^2 = 1$) as its special case. Thus, the results derived based on a nonadditive model in this article generalize the applications of G theory to cases with interactions between subject and facet (e.g., essay and rater), which is referred to as nonadditive G theory.

The formulas for estimated VCs under additive and nonadditive one-facet models are summarized and presented in Table 2. Note that the estimators of the VCs for σ_i^2 and $\sigma_{pi,e}^2$ are the same for both the additive and nonadditive models.

The estimators of VCs for a fixed facet and for a random facet coincide when interactions between subject and facet are absent. However, when such an interaction exists, but an additive model is inadvertently used to analyze the data, $\hat{\sigma}_p^2(\text{additive}) \equiv \hat{\sigma}_p^2(\hat{\lambda}^2 = 1)$ as an estimator of σ_p^2 will underestimate σ_p^2 . Because $\hat{\sigma}_p^2(\lambda^2)$ is an unbiased estimator of σ_p^2 if λ^2 is known, the bias of $\hat{\sigma}_p^2(\text{additive})$ as an estimator of σ_p^2 can be approximated by

$$\hat{\sigma}_p^2(\hat{\lambda}^2 = 1) - \hat{\sigma}_p^2(\hat{\lambda}^2) = - \frac{(1 - \hat{\lambda}^2) MS_{pi,e}}{n_i}$$

Table 2. Estimated VCs in a One-Facet Design.

Sources of variation	Estimated VCs		
	Additive Model 1 with a fixed or random facet	Additive Model 2 with a fixed or random facet	Nonadditive Model 1 with a fixed facet
Person (p)	$\hat{\sigma}_p^2 = \frac{MS_p - MS_{pi,e}}{n_i}$	$\hat{\sigma}_p^2 = \frac{MS_p - MS_{pi,e}}{n_i}$	$\hat{\sigma}_p^2 = \frac{MS_p - \lambda^2 MS_{pi,e}}{n_i}$
Item (i)	$\hat{\sigma}_i^2 = \frac{MS_i - MS_{pi,e}}{n_p}$	$\hat{\sigma}_i^2 = \frac{MS_i - MS_{pi,e}}{n_p}$	$\hat{\sigma}_i^2 = \frac{MS_i - MS_{pi,e}}{n_p}$
pi, e	$\hat{\sigma}_{pi,e}^2 = MS_{pi,e} = \hat{\sigma}_e^2 + \hat{\sigma}_{pi}^2$	$\hat{\sigma}_{pi,e}^2 = MS_{pi,e} = \hat{\sigma}_e^2$	$\hat{\sigma}_{pi,e}^2 = MS_{pi,e}$

Note. VC = variance component; MS = mean squares.

or expressed with population parameters as

$$\text{bias of } \hat{\sigma}_p^2(\text{additive}) = -\frac{(1 - \lambda^2)\sigma_{pi,e}^2}{n_i}. \quad (8)$$

Equation 8 can be used to evaluate the bias if the additivity is assumed as current G theory does. When a data set follows an additive model (i.e., $\lambda^2 = 1$, the largest value), the bias is expected to be zero. When $\lambda^2 < 1$ but the number of facet levels is fairly large, the bias is still present but should be small or even negligible. However, when n_i is small and $\lambda^2 < 1$, the bias can be severe, which may lead to a negative estimate of σ_p^2 , especially when the magnitude of the negative bias is larger than the value of the variance of persons/subjects, $((1 - \lambda^2)\sigma_{pi,e}^2)/n_i > \sigma_p^2$. The above results are summarized as the following theorem.

Theorem 2: (a) The estimator of the VC for subjects given by current G theory is unbiased if, and only if, no genuine interaction exists. (b) In general, the bias of the estimator of the VC for subjects given by current G theory can be approximated by Equation 8. The magnitude of the negative bias is a strictly decreasing function of the additivity index λ^2 and the number of facet levels n_i : the smaller the value of λ^2 or n_i , the larger the magnitude of the bias.

The VCs are the building blocks of many important indexes such as the intraclass correlation coefficient (ICC) in ANOVA and the G coefficients in G theory. Readers can easily find the similarity between the ICC and G coefficients from the formulas presented below.

ICCs are used to measure the strength of association and statistical dependence:

$$ICC_{\text{Fixed}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pi,e}^2} \quad (9)$$

for a fixed facet, and

$$ICC_{\text{Random}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2} \quad (10)$$

for a random facet. For details about how to calculate estimated ICCs for fixed- and random-effect models, see Shrout and Fleiss (1979).

As mentioned in the previous section, an analysis in G theory typically consists of two steps, a generalizability study and a decision study, and a decision can be a relative or an absolute decision. The G coefficient for a relative decision is defined as

$$\phi_{\text{Rel}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{Rel}}^2} \quad (11)$$

where

$$\sigma_{\text{Rel}}^2 = \frac{\sigma_{pi,e}^2}{n'_i}.$$

Here, n'_i refers to the number of levels of the facet used in the decision study. The G coefficient for an absolute decision (also called an *index of dependability*) is defined as

$$\phi_{\text{Abs}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{Abs}}^2}. \quad (12)$$

where

$$\sigma_{\text{Abs}}^2 = \frac{\sigma_i^2}{n'_i} + \frac{\sigma_{pi,e}^2}{n'_i}.$$

When $n'_i = 1$, the formulas for the G coefficient for a relative decision (Equation 11) and for the ICC for a fixed facet (Equation 9) are the same, and the formulas for the index of dependability (Equation 12) and for the ICC for a random facet (Equation 10) are the same.

Notice that $y = x / (x + c)$ is a monotone increasing function of x , when $c > 0$. Thus, if σ_p^2 is underestimated, so are the ICC, ϕ_{Rel} , and ϕ_{Abs} . The estimated G coefficient for a relative decision based on an additive model (when interaction effects are not significant) is

$$\hat{\phi}_{\text{Rel}} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{\text{Rel}}^2}. \quad (13)$$

and the one based on a nonadditive model (when interaction effects are significant) is

$$\hat{\phi}_{\text{Rel}}(\hat{\lambda}^2) = \frac{\hat{\sigma}_p^2(\hat{\lambda}^2)}{\hat{\sigma}_p^2(\hat{\lambda}^2) + \hat{\sigma}_{\text{Rel}}^2}. \quad (14)$$

where

$$\hat{\sigma}_{\text{Rel}}^2 = \frac{\hat{\sigma}_{pi,e}^2}{n'_i}.$$

Similarly, the estimated G coefficients for an absolute decision based on additive and nonadditive models can be obtained. Clearly, the estimated ICC, $\hat{\phi}_{\text{Rel}}$, and $\hat{\phi}_{\text{Abs}}$ will be negative when the estimate of σ_p^2 is negative.

Ideally, the facet in a one-facet design should be treated as random because a G-theory user intends to generalize results to the universe of generalization. Thus, the existence of subject-by-facet interaction effects is not desirable in a generalizability study with one facet. If interaction

effects are significant, depending on the nature of the facet, appropriate actions, such as intensive rater training, item modification, or form reconstruction, may be called for to eliminate possible causes of the interaction in any future studies, including the follow-up decision study. If possible, the generalizability study should be repeated independently; that is, data should be recollected so as to ensure that new data contain no interaction (e.g., essays are rescored by newly trained raters). Clearly, this is an ideal way to deal with situations where the initial data contain significant interaction effects. However, in practice, data recollection is typically not feasible and one has to use the original data set to do any analyses. In such a case, not all selected nominally “qualified raters” (conditions of measurement, in general) are truly qualified (admissible) and thus, the nonadditive model should be applied to isolate the interactions and better estimate the VC for subjects. Regardless of the presence or absence of interaction, it is reasonable to assume that the subject effect and the rater effect based on either nominally “qualified raters” or truly “qualified raters” will be the same. That is, nominally “qualified raters” are expected to produce the same VCs for subjects and raters in a Nonadditive Model (1) as those produced by truly “qualified raters” in a Random-Facet Additive Model 2. As such, the estimated VCs from a Nonadditive Model 1 can be regarded as the corresponding VC estimates for a Random-Facet Additive Model 2, the G coefficients can be estimated by Equation 14, and a decision study can be performed. That is, a nonadditive model is used to better estimate the VC of subjects only when data contain interactions but data recollection is not feasible.

Simulation Studies

Two simulation studies were carried out to evaluate and compare the performance of the nonadditivity approach developed in the preceding section with that of the additive approach used in current G theory. The first study tested the nonadditivity approach on simulated data without any interaction (i.e., data generated from an additive model or $\lambda^2 = 1$) and examined whether the results obtained from these two approaches were at the similar level of accuracy. The second study considered the scenario when interactions existed but the additive approach was used to evaluate the impact of failing to consider interaction effects and to assess the improvement of analysis results by applying the nonadditive approach. Clearly, the settings of the two simulation studies represent two scenarios where additive and nonadditive models are true models, respectively.

Simulation Settings and Data Generation

In both simulation studies, data sets were generated from Equation 1. Items were targeted to be scored as 0, 1, 2, or 3. The overall mean was set to be $\mu = 1.5$. The values of VCs were chosen according to Table 3.2 in Shavelson and Webb (1991) with a certain adjustment for the score scale from 1 to 3, that is, the authors used 9 times of the original values (i.e., $\sigma_p^2 = 0.2745$, $\sigma_i^2 = 0.0837$, and $\sigma_{pi,e}^2 = 1.8927$).

In the first simulation study, the subject effects $\{\tau_p\}$ were randomly drawn from $N(0, \sigma_p^2)$, the facet effects $\{\alpha_i\}$ from $N(0, \sigma_i^2)$, and the error components $\{\varepsilon_{pi}^* = (\tau\alpha)_{pi} + \varepsilon_{pi}\}$ from $N(0, \sigma_{pi,e}^2)$ independently. After all components were generated, $\mathbf{X}_{pi}$ was calculated based on Equation 1 by adding the components together. That is, an additive random-effect model was used to generate data, and, thus, $\lambda^2 = 1$.

In the second study, the subject effects $\{\tau_p\}$ were randomly drawn from $N(0, \sigma_p^2)$, whereas the facet effects $\{\alpha_i\}$ were drawn from a normal distribution with constraints that $\sum_i \alpha_i = 0$ and variance σ_i^2 . The interaction effects $\{(\tau\alpha)_{pi}\}$ and the pure error components $\{\varepsilon_{pi}\}$ were generated separately. Specifically, the interaction and the pure error components equally share $\sigma_{pi,e}^2$,

that is, $\sigma_e^2 = \sigma_{pi}^2 = 0.94635$ and $\lambda^2 = 0.5$. The pure error terms $\{\varepsilon_{pi}\}$ were randomly generated from $N(0, \sigma_e^2)$ and independent from all other terms. The interaction effects $\{(\tau\alpha)_{pi}\}$ were produced by assuming that $(\tau\alpha)_{pi} = \eta\tau_p\alpha_i$, where η is a constant scale factor such that $\{(\tau\alpha)_{pi}\}$ have variance $\sigma_{pi}^2 = 0.94635$. Clearly, $\sum_i (\tau\alpha)_{pi} = 0$ for each p . As such, the interaction effects are correlated with the subject and facet effects. After all terms were generated, a data matrix of $\mathbf{X}_{pi}$ based on Equation 1 with interaction was constructed.

In both studies, the number of persons were selected to be $n_p = 25, 50, 100$ or $1,000$, and the items (or raters) $n_i = 2, 3, 4, 5, 6, 10$, or 20 . Hence, there are 28 combinations (conditions) considered in each of the two simulation studies. One thousand replications were conducted for each combination. Thus, a total of 28,000 sets of data were generated for each simulation study.

Both the additive approach of current G theory and the nonadditive approach developed in this article were applied to analyze each set of simulated data and results were summarized and compared. Note that these two approaches will produce exactly the same estimation results for σ_i^2 and $\sigma_{pi,e}^2$. Thus, when comparing VC estimation results between these two approaches, one only needs to compare the results of estimated VCs for persons, $\hat{\sigma}_p^2$ using the additive approach (i.e., Equation 6) and $\hat{\sigma}_p^2(\hat{\lambda}^2)$ using the nonadditive approach (i.e., Equation 7).

Simulation Study 1: Additive Model ($\lambda^2 = 1$)

In this study, data were generated from Equation 1 with random effects (i.e., an additive model) satisfying all of the assumptions of current G-theory models. The results are summarized in Table 3.

The third column presents the averages of estimated λ^2 based on 1,000 replications for the 28 conditions considered in the article. The average of estimated λ^2 ranges from 0.9678 to 0.9799, which is as expected because the significance level for Tukey's test is 0.05 and the test controls the rate of Type I errors well. The last two columns are the average biases of estimated VCs for persons based on 1,000 replications from the additive and nonadditive approaches, respectively. In some cases, the average biases from one approach are smaller than those from the other, but the opposite is true in other cases. Overall, there is no noticeable difference in bias between these two approaches, which indicates that when applied to data without any interaction, the performance of these two approaches is effectively the same.

Simulation Study 2: Nonadditive Model ($\lambda^2 = 0.5$)

In this study, data were generated from Equation 1 with interaction and $\lambda^2 = 0.5$. In addition to calculating $\hat{\sigma}_p^2$ using the additive approach and $\hat{\sigma}_p^2(\hat{\lambda}^2)$ using the nonadditive approach, one additional VC for persons were calculated for each data set by using Equation 7 with the true $\lambda^2 = 0.5$ instead of $\hat{\lambda}^2$, denoted as $\hat{\sigma}_p^2(\lambda^2)$. In practice, the additivity index λ^2 is an unknown parameter. In simulation, however, the results based on the true λ^2 can be used as a reference to examine the best results one can obtain if λ^2 could be perfectly estimated and thus, to evaluate the effectiveness of Tukey's method in estimating λ^2 . Based on 1,000 replications, the average bias and root mean square error (RMSE) of the estimated VCs for persons, the frequency of negative estimated VCs among 1,000 replications, and the average bias of estimated G coefficients for a relative decision were calculated. The results were summarized in Tables 4 to 7.

Tables 4 and 5 present the average biases and RMSEs of estimated VCs for persons, respectively. The average biases and RMSEs of $\hat{\sigma}_p^2$ are relatively large, especially when the number

Table 3. Average Estimated Additivity Index and Biases of Estimates of the VC for Persons Based on 1,000 Replications When Data Contain No Interaction.

n_i	n_p	$\hat{\lambda}^2$	Bias of estimated VC	
			Additive	Nonadditive
2	25	0.9725	−0.0310	−0.0065
	50	0.9721	−0.0065	0.0180
	100	0.9731	−0.0077	0.0175
	1,000	0.9784	−0.0011	0.0196
3	25	0.9678	−0.0160	0.0058
	50	0.9769	0.0001	0.0145
	100	0.9765	0.0064	0.0214
	1,000	0.9689	−0.0007	0.0190
4	25	0.9708	−0.0012	0.0125
	50	0.9758	0.0052	0.0162
	100	0.9746	−0.0004	0.0117
	1,000	0.9780	0.0007	0.0111
5	25	0.9762	0.0048	0.0140
	50	0.9765	−0.0059	0.0032
	100	0.9778	−0.0045	0.0039
	1,000	0.9755	0.0001	0.0094
6	25	0.9741	0.0091	0.0174
	50	0.9751	−0.0075	0.0004
	100	0.9753	−0.0005	0.0073
	1,000	0.9715	0.0000	0.0090
10	25	0.9719	0.0066	0.0120
	50	0.9743	−0.0057	−0.0009
	100	0.9759	0.0009	0.0055
	1,000	0.9730	0.0003	0.0054
20	25	0.9752	−0.0028	−0.0005
	50	0.9719	−0.0009	0.0018
	100	0.9799	0.0034	0.0053
	1,000	0.9785	0.0002	0.0022

Note. VC = variance component; n_i = number of items; n_p = number of persons.

of facet levels, n_i (e.g., number of raters), is small. According to Theorem 2 or Equation 8, the expected bias in the simulation is

$$-\frac{(1 - \lambda^2)\sigma_{pi,e}^2}{n_i} = -\frac{(1 - 0.5)1.8927}{n_i} = -\frac{1.8927}{2n_i}.$$

Note that the magnitude of the bias from an additive model depends on n_i , but not on n_p . Thus, the biases of $\hat{\sigma}_p^2$ in Table 4 with different numbers of persons are approximately the same when n_i is held constant. For example, when $n_i=2$, the expected bias of $\hat{\sigma}_p^2$ should be around −0.4732, no matter how large the number of persons is. Table 4 show that the estimated average biases of $\hat{\sigma}_p^2$ for 25, 50, 100, and 1,000 persons are −0.4712, −0.4901, −0.4771, and −0.4733, respectively. The corresponding biases of $\hat{\sigma}_p^2(\hat{\lambda}^2)$ from the nonadditive approach for 25, 50, 100, and 1,000 subjects are reduced to −0.2643, −0.1653, −0.0130, and 0.0440, respectively. Obviously, when interactions are present, the average biases are less severe in the estimator based on a nonadditive model. In all cases, the biases of $\hat{\sigma}_p^2(\hat{\lambda}^2)$ are noticeably less than the biases of $\hat{\sigma}_p^2$. However, the biases of $\hat{\sigma}_p^2(\lambda^2)$, where the true additivity index λ^2 is

Table 4. Average Biases of Three Estimates of the VC for Persons Based on 1,000 Replications.

n_i	n_p	Additive	Nonadditive	
			With $\hat{\lambda}^2$	With $\lambda^2 = .5$
2	25	-0.4712	-0.2643	0.0043
	50	-0.4901	-0.1653	-0.0109
	100	-0.4771	-0.0130	-0.0035
	1,000	-0.4733	0.0440	0.0005
3	25	-0.3135	-0.2447	0.0015
	50	-0.3231	-0.1994	-0.0077
	100	-0.3146	-0.1222	-0.0007
	1,000	-0.3159	-0.0227	-0.0003
4	25	-0.2351	-0.1277	0.0002
	50	-0.2360	-0.0794	-0.0009
	100	-0.2384	-0.0558	-0.0010
	1,000	-0.2365	-0.0110	-0.0003
5	25	-0.1876	-0.1152	0.0022
	50	-0.1879	-0.0779	0.0026
	100	-0.1899	-0.0652	-0.0003
	1,000	-0.1886	-0.0173	0.0009
6	25	-0.1626	-0.1131	-0.0054
	50	-0.1592	-0.0850	-0.0013
	100	-0.1596	-0.0665	-0.0019
	1,000	-0.1573	-0.0190	0.0004
10	25	-0.0974	-0.0526	-0.0023
	50	-0.0938	-0.0366	0.0011
	100	-0.0945	-0.0335	0.0003
	1,000	-0.0948	-0.0104	-0.0002
20	25	-0.0380	-0.0146	0.0096
	50	-0.0451	-0.0184	0.0023
	100	-0.0467	-0.0153	0.0007
	1,000	-0.0482	-0.0066	-0.0009

Note. VC = variance component; n_i = number of items; n_p = number of persons.

assumed to be known (but the VCs are estimated from data), are near zero, and much smaller than the biases of $\hat{\sigma}_p^2(\hat{\lambda}^2)$ (compare the last two columns of Table 4), especially when the sample size is relatively small. This observation indicates that when the sample size is relatively small, the estimation of λ^2 is not accurate enough; that is, it requires a large number of subjects/persons to get an accurate estimate of λ^2 based on Tukey's method. The results in Tables 5 and 6 will also confirm this conclusion.

Table 5 shows that the RMSEs of $\hat{\sigma}_p^2(\hat{\lambda}^2)$ are uniformly smaller than the RMSEs of $\hat{\sigma}_p^2$, but larger than the RMSEs of $\hat{\sigma}_p^2(\lambda^2)$, across all cases considered in the simulation except for $n_i = 20$ and $n_p = 25$. For example, the RMSE of $\hat{\sigma}_p^2(\hat{\lambda}^2)$ is 0.2033, whereas the corresponding ones for $\hat{\sigma}_p^2$ and $\hat{\sigma}_p^2(\lambda^2)$ are 0.5033 and 0.1189, respectively, when $n_i = 2$ and $n_p = 100$.

Table 6 shows that the frequency of getting negative $\hat{\sigma}_p^2(\hat{\lambda}^2)$ is uniformly lower than that of getting negative $\hat{\sigma}_p^2$ except for the cases where there is no negative $\hat{\sigma}_p^2$. For instance, out of 1,000 replications, the counts of getting negative $\hat{\sigma}_p^2$ for $n_i = 2$ are 725, 822, 911, and 1,000 for 25-, 50-, 100-, and 1,000-person cases, respectively, while the corresponding counts for $\hat{\sigma}_p^2(\hat{\lambda}^2)$ are 487, 353, 104, and 0. When the number of subjects/persons is relatively small for a given n_i in the simulation, the frequencies of getting negative $\hat{\sigma}_p^2(\hat{\lambda}^2)$ are much larger than the corresponding ones for $\hat{\sigma}_p^2(\lambda^2)$ where the additivity index λ^2 is regarded as known. Note that it is possible to

Table 5. Average RMSEs of Three Estimates of the VC for Persons Based on 1,000 Replications.

n_i	n_p	Additive	Nonadditive	
			With $\hat{\lambda}^2$	With $\lambda^2 = .5$
2	25	0.5777	0.4851	0.2436
	50	0.5413	0.3749	0.1669
	100	0.5033	0.2033	0.1189
	1,000	0.4761	0.0584	0.0378
3	25	0.3705	0.3480	0.1720
	50	0.3553	0.2958	0.1279
	100	0.3305	0.2267	0.0864
	1,000	0.3176	0.1049	0.0280
4	25	0.2820	0.2412	0.1422
	50	0.2592	0.1901	0.1023
	100	0.2505	0.1479	0.0711
	1,000	0.2377	0.0747	0.0216
5	25	0.2326	0.2131	0.1335
	50	0.2100	0.1613	0.0904
	100	0.2010	0.1351	0.0640
	1,000	0.1898	0.0690	0.0212
6	25	0.2033	0.1893	0.1206
	50	0.1827	0.1507	0.0877
	100	0.1711	0.1208	0.0612
	1,000	0.1584	0.0628	0.0189
10	25	0.1426	0.1320	0.1059
	50	0.1168	0.0974	0.0712
	100	0.1079	0.0791	0.0529
	1,000	0.0961	0.0381	0.0161
20	25	0.0986	0.0999	0.0927
	50	0.0760	0.0706	0.0623
	100	0.0642	0.0541	0.0447
	1,000	0.0502	0.0221	0.0142

Note. RMSE = root mean square error; VC = variance component; n_i = number of items; n_p = number of persons.

get negative VC estimates even when a true model, either additive or nonadditive, is used, but it should not happen too frequently (see Arnold, 1981).

The estimated G coefficients for a relative decision, $\hat{\phi}_{\text{Rel}}$, $\hat{\phi}_{\text{Rel}}(\hat{\lambda}^2)$, and $\hat{\phi}_{\text{Rel}}(\lambda^2 = 0.5)$, were calculated. Specifically, $\hat{\phi}_{\text{Rel}}$ was calculated using the additive approach (Equation 13), and $\hat{\phi}_{\text{Rel}}(\hat{\lambda}^2)$ and $\hat{\phi}_{\text{Rel}}(\lambda^2 = 0.5)$ were calculated using the nonadditive approach (Equation 14). Table 7 presents the average biases of these estimated G coefficients for a relative decision based on 1,000 replications. The magnitudes of the average biases of $\hat{\phi}_{\text{Rel}}$ are rather large, especially in the cases where $n_i < 10$ (see the third column of Table 7). When the nonadditive approach was applied, the magnitudes of the average biases of estimated G coefficients (the fourth column of Table 7) become noticeably smaller than the corresponding ones (the third column) from the additive approach, but still substantially larger than the corresponding ones from the nonadditive approach with the known true λ^2 (the last column of Table 7), which indicates that the nonadditive approach can improve the accuracy of G coefficient estimation, but still has room for improvement if λ^2 can be estimated better than using Tukey's method.

In sum, the second simulation study confirms the theoretical results obtained in the preceding section that when data contain the subject-by-facet interaction, the VC estimator for subjects/persons based on the formula from the current G theory (i.e., additive approach) has a

Table 6. Frequency of Negative Estimated VCs for Persons Among 1,000 Replications.

n_i	n_p	Additive	Nonadditive	
			With $\hat{\lambda}^2$	With $\lambda^2 = .5$
2	25	725	487	116
	50	822	353	52
	100	911	104	11
	1,000	1,000	0	0
3	25	589	485	42
	50	648	414	6
	100	670	283	0
	1,000	894	86	0
4	25	404	251	9
	50	390	174	1
	100	330	100	0
	1,000	43	2	0
5	25	273	195	7
	50	172	89	0
	100	92	31	0
	1,000	0	0	0
6	25	169	131	3
	50	85	48	0
	100	20	8	0
	1,000	0	0	0
10	25	18	11	0
	50	0	0	0
	100	0	0	0
	1,000	0	0	0
20	25	0	0	0
	50	0	0	0
	100	0	0	0
	1,000	0	0	0

Note. VC = variance component; n_i = number of items; n_p = number of persons.

substantial negative bias when the number of facet levels is small. The nonadditive approach with Tukey’s test for nonadditivity can reduce the bias substantially in such cases. Although the simulation study shows that the estimator based on the nonadditive approach works well in terms of reducing bias and RMSE, there are still many cases with a relatively large bias of $\hat{\sigma}_p^2(\hat{\lambda}^2)$ compared with the bias of $\hat{\sigma}_p^2(\lambda^2)$. The reason could be that the estimator of the additivity index based on Tukey’s method is still not accurate enough, especially when the number of subjects is not large (e.g., less than 100). Further discussion is presented in the last section.

Note that the bias of the estimated VCs based on additive models will become small when n_i is relatively large according to Equation 8. In the simulation setting, when n_i is larger than 10, the bias of $\hat{\sigma}_p^2(\text{additive})$ becomes negligible. Thus, when n_i is large enough such that the expected bias given by Equation 8 is small enough, one may safely use G theory with additive models in data analyses regardless of whether interactions exist or not. Nevertheless, in a typical rater-mediated assessment, the number of raters does not usually exceed 10 and, therefore, the bias of estimated VCs based on additive models with a small number of raters warrants practical concerns.

Table 7. Average Biases of Estimated G Coefficients for a Relative Decision Based on 1,000 Replications.

n_i	n_p	Additive	Nonadditive	
			With $\hat{\lambda}^2$	With $\lambda^2 = .5$
2	25	−0.5869	−0.3728	−0.0052
	50	−0.5634	−0.2242	−0.0124
	100	−0.5205	−0.0377	−0.0045
	1,000	−0.4925	0.0265	−0.0001
3	25	−0.4548	−0.3747	−0.0148
	50	−0.4329	−0.2780	−0.0151
	100	−0.3919	−0.1618	−0.0036
	1,000	−0.3756	−0.0314	−0.0008
4	25	−0.3576	−0.2147	−0.0152
	50	−0.3259	−0.1281	−0.0086
	100	−0.3129	−0.0832	−0.0060
	1,000	−0.2942	−0.0178	−0.0003
5	25	−0.3034	−0.2083	−0.0197
	50	−0.2650	−0.1231	−0.0085
	100	−0.2526	−0.0937	−0.0055
	1,000	−0.2370	−0.0244	0.0000
6	25	−0.2621	−0.1967	−0.0244
	50	−0.2272	−0.1300	−0.0120
	100	−0.2125	−0.0918	−0.0069
	1,000	−0.1956	−0.0253	−0.0001
10	25	−0.1509	−0.0951	−0.0228
	50	−0.1226	−0.0568	−0.0087
	100	−0.1148	−0.0448	−0.0050
	1,000	−0.1058	−0.0125	−0.0006
20	25	−0.0522	−0.0329	−0.0087
	50	−0.0461	−0.0238	−0.0053
	100	−0.0427	−0.0171	−0.0033
	1,000	−0.0390	−0.0055	−0.0009

Note. n_i = number of items; n_p = number of persons.

A Real Operation Data Analysis

The data set was collected in summer 2009 during a standards-to-standards correspondence study in Mississippi (Chi, Lin, & Yap, 2009). One of the objectives of the study was to compare cognitive complexity between two sets of standards (descriptors). One set of 25 standards was treated as the anchor standards. The anchor standards are English language proficiency standards designed for English language learners (ELLs). These anchor standards describe linguistic skills and language demands in relation to ELLs being able to understand and convey content knowledge in mainstream classrooms. A panel of six reviewers ($n_i = 6$) rated each anchor standard independently based on an established cognitive scale from 1 to 4. Level 1 represents low cognitive-demand processing, such as simple recall of facts, whereas Level 4 involves higher cognitive complexity, such as synthesizing difference sources of information. The reviewers were school teachers who worked directly with the ELL populations and had experience implementing the anchor standards in their classrooms. Prior to the study, the reviewers attended a workshop and received intensive training on judging the cognitive complexity of the standards.

The standards and reviewers were crossed in the study design; that is, each reviewer rated the same set of 25 anchor standards. Here, these anchor standards were treated as subjects ($n_p = 25$).

Before comparing the cognitive levels of the two sets of standards, Chi et al. (2009) examined the degree to which ratings from the panel of reviewers were reliable. The G coefficients for relative and absolute decisions were used to assess the reliability of the panel of reviewers. For the panel of six reviewers in the pre-Kindergarten to Kindergarten grade cluster, $MS_p = 0.182$, $MS_i = 2.332$, and $S_{pi,e} = 0.204$. Using the regular additive G-theory formulas, one obtains $\hat{\sigma}_p^2 = -0.004$, $\hat{\sigma}_i^2 = 0.085$, and $\hat{\sigma}_{pi,e}^2 = 0.204$. The G coefficients of a panel of six reviewers for relative and absolute decisions were -0.121 and -0.082 , respectively. Clearly, the negative estimated VC leads to the negative G coefficients. In practice, these negative values are typically reported as zeros.

The nonadditive approach was applied to reanalyze the data. The estimate of the additivity index based on Tukey's method is 0.4988. According to Equation 7, $\hat{\sigma}_p^2(\hat{\lambda}^2) = 0.013$, and the values for the other two estimated VCs remain the same. Thus, the new G coefficients are 0.282 and 0.217 for relative and absolute decisions, respectively. Now both G coefficients are positive.

Discussion

This article first discusses the distinction between the fixed and random facets and the distinction between genuine and removable interactions in G theory. Then, it introduces Scheffe's theorem to point out the fact that the random-effect linear G-theory model is actually assumed to be additive, which is equivalent to assuming that no subject-by-facet interaction exists in a one-facet design. Ideally, the facet in a one-facet design should be treated as random and subject-by-facet interactions should be avoided. However, in practice, additivity (no interaction) may be untenable and can be severely violated. In such a case, the VC estimator for subjects based on a random facet or an additive model is expected to have negative bias. As shown by Equation 8, the bias can be severe when the number of selected facet levels is small. Thus, prior to treating the facet as random in statistical analyses, it is important to examine whether interaction effects are significant. Finally, the article expands G theory to a more general setting in which both additive and nonadditive models are considered. Such research efforts are important in that interactions are likely to occur in operational settings, and when they do, applications of nonadditive G theory become necessary. In sum, when interactions are absent or removable, current G-theory methods will suffice, but when interactions exist, nonadditive G theory is needed to estimate VCs.

Given a data set, the nonadditive G-theory approach developed in this article starts with a mixed-effect model to deal with possible interaction effects. First, it uses Tukey's test for nonadditivity to examine whether the subject-by-facet interaction exists. If the test result is significant, then the nonadditive approach estimates the additivity index based on Tukey's method and estimates the VC for subjects using Equation 7; otherwise, it sets the additivity index to be one and uses an additive model or Equation 6 to estimate the VC for subjects just as current G theory does. After obtaining estimated VCs, the rest of the analysis remains the same as that in current G theory. Thus, the nonadditive approach generalizes current G theory to deal with the complication of VC estimates due to nonadditivity in data by adding the initial step of testing the subject-by-facet interaction.

In relation to the real data analysis in the previous section, the practical implications of nonadditive G theory come in two folds. First, the notion of interaction/nonadditivity gives another possible explanation for negative VC estimates in practice. Additive G-theory models are shown to underestimate the VC for subjects in the presence of interaction in data. If the true VC is

expected to be small, the estimated VC could be negative depending on the degree of such underestimation. Second, the use of the additivity index can serve as a quick tool to evaluate the quality of data. A small value in the additivity index would signal the presence of interaction between the subject (e.g., person or essay) and the facet (e.g., item or rater).

In general, it is important to check the nonadditivity (interaction) of any given data. Tukey (1949) investigated this issue and proposed a test for nonadditivity. Tukey decomposed the residual sum of squares into an interaction component representing nonadditivity and the remaining component as an error term for testing nonadditivity. Although Tukey's test was successful in singling out the interaction effects in the second simulation study, it is not sensitive to all interactions. As Kirk (2013) pointed out, "Tukey's procedure tests the (null) hypothesis that a special type of contrast–contrast interaction is equal to zero. For this contrast–contrast interaction, the coefficients of the contrast are the $[\bar{X}_{p.} - \bar{X}_{..}]$ and $[\bar{X}_{.i} - \bar{X}_{..}]$ deviations" (p. 300). Our simulation results show that Tukey's method works well in estimating the additivity index λ^2 but still has room for improvement. One may surmise that similar statistical techniques can also be developed for other types of interactions and, consequently, one can obtain a better estimator of the additivity index in nonadditive G theory or develop a more powerful test for nonadditivity than Tukey's test. Furthermore, Tukey's method also needs to be generalized to designs with multiple facets. It is hoped that this article can spark further discussion and research on ways to deal with interaction/nonadditivity in G-theory applications.

Acknowledgment

The authors would like to thank Sarah Zhang, Xiaohong Gao, and four anonymous reviewers for their comments and suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Arnold, S. F. (1981). *The theory of linear models and multivariate analysis*. New York, NY: John Wiley.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
- Chi, Y., Lin, C., & Yap, J. (2009). *Alignment between the Mississippi curriculum framework in language arts, mathematics, and science and the WIDA consortium English language proficiency standards*. Champaign: University of Illinois at Urbana–Champaign.
- Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.
- 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: 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.
- Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York, NY: Springer.
- Everitt, B. S. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, 44, 113-135.

- Gelman, A. (2005). Analysis of variance—Why it is more important than ever. *Annals of Statistics*, 33, 1-33.
- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.
- Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London, England: Sage.
- Lin, C. (2014). *Issues and challenges in current generalizability theory applications in rated measurement* (Unpublished doctoral dissertation). Department of Educational Psychology, University of Illinois at Urbana-Champaign.
- Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston, MA: Allyn & Bacon.
- Myers, J. L., Well, A. D., & Lorch, R. F. (2010). *Research design and statistical analysis* (3rd ed.). New York, NY: Routledge.
- Scheffe, H. (1959). *The analysis of variance*. New York, NY: John Wiley.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Snijder, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction basic and advanced multilevel modeling* (2nd ed.). London, England: Sage.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Belmont, CA: Thomson Brook/Cole.
- Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, 5, 232-242.