JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
*Journal of Educational Measurement*
*Spring 2018, Vol. 55, No. 1, pp. 78–89*

# Detecting Nonadditivity in Single-Facet Generalizability Theory Applications: Tukey's Test

**Chih-Kai Lin**
*American Institutes for Research*
**Jinming Zhang**
*University of Illinois at Urbana-Champaign*

*Under the generalizability-theory (G-theory) framework, the estimation precision of variance components (VCs) is of significant importance in that they serve as the foundation of estimating reliability. Zhang and Lin advanced the discussion of nonadditivity in data from a theoretical perspective and showed the adverse effects of nonadditivity on the estimation precision of VCs in 2016. Contributing to this line of research, the current article directs the discussion of nonadditivity from a theoretical perspective to a practical application and highlights the importance of detecting nonadditivity in G-theory applications. To this end, Tukey's test for nonadditivity is the only method to date that is appropriate for the typical single-facet G-theory design, in which a single observation is made per element within a facet. The current article evaluates the Type I and Type II error rates of Tukey's test. Results show that Tukey's test is satisfactory in controlling for falsely detecting nonadditivity when the data are actually additive and that it is generally powerful in detecting nonadditivity when it exists. Finally, the article demonstrates an application of Tukey's test in detecting nonadditivity in a judgmental study of educational standards and shows how Tukey's test results can be used to correct imprecision in the estimated VC in the presence of nonadditivity.*

Generalizability theory or G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) conceptualizes observed measurement variability as a combination of the true variation in the objects of measurement, other measurement variation(s) that are anticipated by or of interest to an investigator, and random error. For example, in an essay exam on some scientific knowledge for a group of students, the object of measurement is students' knowledge in science, and a potential source of measurement variation (i.e., *facet* in G-theory terminology) is score variability introduced by different raters scoring the essays. Ideally, one would like to see true differences in students' scientific knowledge reflect observed score variability as much as possible, not differences among rater severity/leniency. In addition to gauging how much observed measurement variability is explicable by different measurement facets, G theory has been widely used in the analysis of measurement reliability in large-scale assessment contexts (Brennan, 2000; Brennan, Gao, & Colton, 1995; Gao, Shavelson, & Baxter, 1994; Lee & Kantor, 2007; Shavelson, Baxter, & Gao, 1993) and in classroom-assessment contexts (Gebril, 2009; Huang & Foote, 2010; Sudweeks, Reeve, & Bradshaw, 2004).

It is noteworthy that in many applications of G theory, the objects of measurement are persons (e.g., student performance on an essay exam); nevertheless, there

 

were many other G-theory applications where the measurement was made on objects other than persons, such as test items and performance standards. For example, in a standard setting study, test-centered procedures may be used to determine cut scores based on raters' judgments on each item in a test. Researchers have used G theory in this context to assess the relative contributions of measurement facet(s) to the variability of raters' judgments on test items (Brennan & Lockwood, 1980; Clauser, Margolis, & Clauser, 2014; Clauser et al., 2008; Yin & Sconing, 2008). Another example is research motivated in part by the mandate, under the No Child Left Behind Act (2002), to conduct content alignment studies for English language learners (ELLs) in the U.S. K-12 setting. Researchers in this context (Lin & Zhang, 2014) have used G theory to investigate the reliability of raters' ratings of the cognitive complexity associated a set of language performance standards; that is, the performance standards were the objects of measurement.

In this article, we purposefully direct our discussion of G-theory applications to cases where the objects of measurement are performance standards, because situating the discussion in this way is consistent with our empirical study design (i.e., ELL content alignment research) presented later in the article. However, the utility of results in this article is not limited to performance standards as the objects of measurement; rather, the results are applicable to any types of objects of measurement under the G-theory framework. Regardless of the types of objects of measurement, studies with G-theory applications are concerned with the replicability or generalizability of results from measurement procedures.

## Single-Facet Design in G Theory

In a G-theory design, an observed measurement is a linear function of the main and interaction effects of standards, facet(s), and errors. For example, a single-facet ($s \times r$) design takes the following linear model:

$$X_{sr} = \mu + \alpha_s + \beta_r + \varepsilon_{sr,e}, \tag{1}$$

where the rating (e.g., observed measurement on some rating scale) ($X_{sr}$) of standard $s$ given by rater $r$ is the sum of an overall mean rating ($\mu$) and the three random effects pertaining to standards, raters, and errors, where $\alpha_s \sim N(0, \sigma_s^2)$, $\beta_r \sim N(0, \sigma_r^2)$, and $\varepsilon_{sr,e} \sim N(0, \sigma_e^2)$, respectively. The overall mean rating can be considered as the average rating of all the objects of measurement (i.e., standards) on the rating scale. The standard effect ($\alpha_s$) reflects the relative rating standing of a standard compared to the overall mean rating, while the rater effect ($\beta_r$) corresponds to the relative severity of a random rater compared to a rater with average rating severity from the *universe of admissible raters*.

In a typical rater-mediated measurement, each object of measurement (e.g., person or standard) is repeatedly rated across some or all raters and is rated once per rater; as a result, the object-of-measurement-by-rater interaction is confounded with the random error, and both the interaction and random-error terms are subsumed under the error component ($\varepsilon_{sr,e}$). It is data with this single-observation-per-cell layout that is typical of G-theory applications. Current G theory assumes additivity such that "all effects in the model are uncorrelated" (Brennan, 2001, p. 23). An example of the

additive assumption is that the three effects of standards, raters, and errors in Model (1) are not correlated with one another.

As will be shown later, when the assumption of additivity is met such that all the effects are uncorrelated, the two confounding terms (i.e., interaction and random error) within the error component do not need to be assessed separately in the estimation of variance components. However, when some or all of the three random effects in Model (1) are correlated, the model becomes nonadditive, and the confounding nature of the error component then introduces additional complications in estimating variance components because the random-error term now needs to be estimated independently of the interaction term (Zhang & Lin, 2016). In both the additive and nonadditive models, the random-error component is uncorrelated with or independent of the other components. The difference between the additive and nonadditive models is that the interaction component is uncorrelated with any components in the additive model, whereas it is correlated with the objects of measurement, the rater effect, or both in the nonadditive model.

Statistically, the distinction between additive and nonadditive models is of great importance in that formulas for variance component estimates differ depending on the nature of the models (see Scheffe, 1999). However, the difference between additivity and nonadditivity is not adequately discussed in the G-theory literature, and hence the identification of potential nonadditivity in data is typically overlooked in G-theory applications. This article attempts to advance the discussion of nonadditivity in G theory and more importantly highlights the importance of detecting nonadditivity in G-theory applications.

## Nonadditivity

The discussion of nonadditivity has been noted in the literature of analysis of variance (ANOVA) regarding data with a single-observation-per-cell design (Davis, 2002; Myers, 1979; Scheffe, 1999). Myers (1979) alluded to the fact that accurate estimation of variance components cannot be achieved with the presence of nonadditivity. In view of the advantages of working with additive data, Anscombe and Tukey (1963) proposed procedures that transform nonadditive data sets into additive ones. As a caveat of preliminary data scrutiny, Scheffe (1999) suggested that it would be helpful to examine the observed variance component for errors, such that a relatively large value may suggest nonadditivity and/or violations of other ANOVA assumption(s). Zhang and Lin (2016) introduced an additivity index, measuring the degree to which additivity assumption holds in data. The larger the index is, the larger the magnitude of additivity would be, and hence the smaller the degree of nonadditivity. By incorporating this index into the G-theory framework, the authors also developed nonadditive G theory for one-facet measurement designs.

Given that the use of G theory leans heavily on ANOVA techniques in estimating variance components and that data associated with G-theory applications usually follow the single-observation-per-cell layout, potential issues with nonadditivity should deserve more attention from G-theory users. In relation to Model (1), Table 1 illustrates the difference between additivity and nonadditivity by presenting formulas for estimated variance components in a one-facet additive G-theory model

Table 1
*Estimated Variance Components for One-Facet ($s \times r$) Additive and Nonadditive Models*

| | Additive Model | Nonadditive Model |
|---|---|---|
| Standard ($s$) | $\hat{\sigma}_s^2 = \frac{\mathrm{MS}_s - \hat{\sigma}_{sr}^2 - \hat{\sigma}_e^2}{n_r}$ | $\hat{\sigma}_s^2 = \frac{\mathrm{MS}_s - \hat{\sigma}_e^2}{n_r}$ |
| Rater ($r$) | $\hat{\sigma}_r^2 = \frac{\mathrm{MS}_r - \hat{\sigma}_{sr}^2 - \hat{\sigma}_e^2}{n_s}$ | $\hat{\sigma}_r^2 = \frac{\mathrm{MS}_r - \hat{\sigma}_{sr}^2 - \hat{\sigma}_e^2}{n_s}$ |
| Error ($sr$, $e$) | $\hat{\sigma}_{sr,e}^2 = \mathrm{MS}_{sr,e} = \hat{\sigma}_{sr}^2 + \hat{\sigma}_e^2$ | $\hat{\sigma}_{sr,e}^2 = \mathrm{MS}_{sr,e} = \hat{\sigma}_{sr}^2 + \hat{\sigma}_e^2$ |

*Note.* MS refers to observed mean squares.

(see Shavelson & Webb, 1991, pp. 28, 29) and for those in a one-facet nonadditive G-theory model (Scheffe, 1999).

The difference between additive and nonadditive assumptions has implications for estimating the variance component for the objects of measurement (i.e., $\hat{\sigma}_s^2$). As such, in the numerator of $\hat{\sigma}_s^2$ in Table 1 for the additive model, the $\mathrm{MS}_s$ is subtracted by both the interaction ($\hat{\sigma}_{sr}^2$) and random error ($\hat{\sigma}_e^2$) components. On the other hand, in the numerator of $\hat{\sigma}_s^2$ for the nonadditive model, the $\mathrm{MS}_s$ is subtracted by only the random-error term ($\hat{\sigma}_e^2$) (more discussion on the estimation of the random-error term is presented in the Method section). Consequently, if an additive model is inadvertently used to analyze nonadditive data, the estimated variance component for the objects of measurement can be adversely underestimated. Zhang and Lin (2016) have shown that when the magnitude of the interaction effect is substantial in a single-facet design, the variance component for the objects of measurement can be unduly underestimated, thereby affecting the estimation of phi-coefficients (analogous to reliability coefficients in classical test theory) in the G-theory framework. They have further shown that in some cases, the underestimation can result in negative variance estimates, which is against the concept of a variance component (see Brennan, 2001, for more detail on negative variance estimates). On the other hand, when the interaction is insignificantly small or does not exist (i.e., $\sigma_{sr}^2 = 0$), the additive and nonadditive models do not differ with respect to variance component estimates. The current G theory requires data be additive in a one-facet design. However, nonadditivity can appear in practice.

## Purpose of the Study

In light of the adverse effects of nonadditivity in data on the estimated variance component for the objects of measurement and consequently on the estimated phi-coefficient (Zhang & Lin, 2016), the current study contributes to this line of research by exploring statistical hypothesis testing that can detect nonadditivity in data under the G-theory framework. Successful identification of nonadditivity is of significant importance in G-theory applications so that the variance component for the objects of measurement can be better estimated. Tukey's single-degree-freedom test for nonadditivity (Tukey, 1949) is the only method to date that has been developed for the single-observation-per-cell type of measurement in testing the significance (or lack thereof) of nonadditivity in data. What has not been investigated in the literature is the usefulness of Tukey's single-degree-freedom test in detecting nonadditivity in

data. In view of the impact of nonadditivity on G-theory analysis, the purposes of the current study are to:

- evaluate Tukey's single-degree-freedom test for nonadditivity in terms of Type I and Type II error rates under the G-theory framework; and
- demonstrate the application of Tukey's test to an empirical data set and highlight its usefulness in correcting for the underestimation of variance component when it occurs.

## Method

The logic behind Tukey's test is briefly sketched here, and readers are directed to Tukey (1949) for more detail on the statistical procedures. First, Tukey's test isolates the sum of squares of a single-degree-freedom nonadditive interaction contrast from the sum of squares of the confounding error component ($\varepsilon_{sr,e}$). Second, it performs a hypothesis test (i.e., $H_o$: $\sigma^2_{sr} = 0$, $H_1$: $\sigma^2_{sr} \neq 0$) regarding the nonadditive interaction contrast via an $F$ ratio statistic:

$$F_{\text{Tukey}} = \frac{\text{SS}_{sr}/1}{(\text{SS}_{sr,e} - \text{SS}_{sr})/(df_{sr,e} - 1)}, \tag{2}$$

where $\text{SS}_{sr}$ is the observed sum of squares of the nonadditive interaction contrast, $\text{SS}_{sr,e}$ is the observed sum of squares of the error component, and $df_{sr,e}$ is the degree of freedom associated with $\text{SS}_{sr,e}$. The observed $F$ ratio is to be compared with $F_{.05}(1, df_{sr,e} - 1)$. A lack of significance for the interaction contrast would lend support to additivity (i.e., $H_o$), while a significant interaction contrast points to nonadditivity (i.e., $H_1$).

When Tukey's test indicates significant nonadditivity in data, one should use the $\sigma^2_s$ under the nonadditive model (i.e., the nonadditive variance component for the objects of measurement) in Table 1. To obtain $\hat{\sigma}^2_s$, one first needs to estimate $\sigma^2_e$ via the partial omega squared for the nonadditive interaction contrast ($\omega^2_{(sr)}$). The definition of $\omega^2_{(sr)}$ (see Keppel & Wickens, 2004, p. 165) may be presented as $\omega^2_{(sr)} = \sigma^2_{sr}/\sigma^2_{sr,e}$, which equals to 1 minus the additivity index introduced by Zhang and Lin (2016) and can be regarded as the nonadditivity index.

The variance component for errors is the composite of the nonadditive interaction and the random error: $\sigma^2_{sr,e} = \sigma^2_{sr} + \sigma^2_e$. Hence, for the estimated nonadditive variance component for the objects of measurement from Table 1, where $\hat{\sigma}^2_s = (\text{MS}_s - \hat{\sigma}^2_e)/n_r$, the component $\hat{\sigma}^2_e$ becomes $\hat{\sigma}^2_{sr,e}(1 - \hat{\omega}^2_{(sr)})$. Next, one would estimate the partial omega squared based on the observed Tukey test's $F$ ratio for the interaction contrast ($F_{\text{Tukey}}$) from Equation 2 as follows:

$$\hat{\omega}^2_{(sr)} = \frac{(F_{\text{Tukey}} - 1)}{(F_{\text{Tukey}} - 1 + 2n_s)}.$$

Finally, the $\hat{\sigma}^2_s$ for the nonadditive model then becomes

$$\hat{\sigma}^2_s = \frac{\text{MS}_s - \hat{\sigma}^2_{sr,e}\left(1 - \hat{\omega}^2_{(sr)}\right)}{n_r}.$$

In the current study, we evaluated Tukey's test in terms of Type I and Type II error rates under the G-theory framework via a simulation study; in addition, we applied Tukey's test to an empirical judgmental study of educational standards, in which a panel of raters rated the cognitive complexity of a set of English language performance standards by using an established rating scale. Upon finding significant nonadditivity in data based on Tukey's test, we also estimated the nonadditive variance component for the objects of measurement. We used the R statistical software, version 2.15.2, to perform the analysis in this study. Next, we provide details on the simulation and empirical studies.

## Simulation Study

Type I error rate is defined as the chance of falsely rejecting a null hypothesis when in fact it is true (Howell, 2013). Generally speaking, for a statistical test to be considered useful, the Type I error rate should be small and is usually set at .05 as a rule of thumb. By situating the notion of Type I error rate in the current study, it translates to the probability of Tukey's test in showing erroneous significant interaction effects for nonadditivity when the data are actually additive. Given that, data generation for the purpose of evaluating the Type I error of Tukey's test followed the assumption of additivity, such that the three random effects in Model (1) were generated independently from three normal distributions, where $\alpha_s \sim N(0, .0305)$, $\beta_r \sim N(0, .0093)$, and $\varepsilon_{sr,e} \sim N(0, .2103)$, respectively. The values of variance components were taken from Shavelson and Webb (1991, p. 29). By generating the three random components independently of one another, one can be certain that nonadditivity is not present because nonadditivity occurs only when some or all of the components are correlated.

Type II error rate is defined as the chance of failing to reject the null hypothesis given that the null hypothesis is actually false (Howell, 2013). What might be more intuitive in the discussion of Type II error rate is the notion of statistical power. Power is defined as the probability of accurately rejecting the null hypothesis when in fact it is false. High power for a statistical test is desirable, and satisfactory power is usually set at .80. Applying the notion of statistical power to Tukey's test would indicate its ability to accurately detect significant nonadditive interaction when the data are in fact nonadditive.

### Simulation Designs

In both the Type I error analysis and the power analysis of Tukey's test, we included four sample sizes ($n_s$): 25, 50, 100, and 1,000 and four numbers of raters ($n_r$): 3, 5, 10, and 20; therefore, we considered a total of 16 conditions in the simulation study. Myers (1979) argued that Tukey's test was particularly sensitive to "correlation between a subject's average performance and the rate at which his performance changes relative to the changes in the group performance" (p. 185). In the current study, this correlation suggests that for objects of measurement (i.e., standards) that are truly high on the rating scale, lenient raters are likely to award higher ratings while harsh raters tend to be more conservative in their ratings. For objects of measurement that are truly low on the rating scale, lenient raters would not uniformly give higher ratings; likewise, harsh raters would not necessarily assign lower ratings.

Due to the different rating patterns in relation to where the objects of measurement stand on the rating scale, a significant object-of-measurement-by-rater interaction exists and thereby constitutes nonadditivity.

Data generation for the purpose of evaluating the statistical power of Tukey's test aimed to incorporate the above correlation identified by Myers (1979). In a $s \times r$ data matrix such as that used in the current study, where standards constitute the rows and raters the columns, this correlation is operationalized as the correlation between the average ratings of standards across all raters ($\bar{X}_{s.}$) and the sum of cross-products of the rating of each standard and the deviation of average rater rating from the overall mean rating ($\sum_r X_{sr}(\bar{X}_{.r} - \bar{X}_{..})$). In the current simulation study, we targeted this correlation at .50 so that it represents a medium magnitude of nonadditivity. The actual average correlation was .54 across all the simulated conditions. The correlation was realized by adding an interaction effect ($\alpha\beta_{sr} \sim N(0, \sigma^2_{sr})$ in Model (1) so that the interaction correlated with both the objects of measurement ($\alpha_s$) and the rater effect ($\beta_r$), while the random error remains to be uncorrelated with any effects. By allowing the objects of measurement and the rater effect to be correlated with the interaction effect, the simulated data become nonadditive.

## Results

For each simulated condition, we conducted 1,000 replications. Hence, Type I error is calculated as the number of replications out of 1,000 in which Tukey's test erroneously suggests the presence of nonadditivity, whereas power is calculated as the number of replications out of 1,000 in which Tukey's test is successful in detecting nonadditivity in the data. Table 2 presents results of Type I error rates for Tukey's test across the 16 simulated conditions.

Results show that the Type I error rate of Tukey's test is around .05 for each condition, suggesting that the test is successful in keeping the occurrences of falsely detecting nonadditivity low when the data are actually additive. Table 3 shows the results of power analysis for Tukey's test.

As expected, when the number of objects of measurement ($n_s$) is fixed, the power increases as the number of raters increases. For example, when $n_s = 50$, the power increases from .69 to 1.00 as the number of raters increases from 3 to 20. In a similar vein, when the number of raters is fixed, the power improves with more objects of

Table 2
*Type I Error of Tukey's Test: False Detection of Nonadditivity in One-Facet Design (1,000 Replications per Condition)*

| | $n_s = 25$ | $n_s = 50$ | $n_s = 100$ | $n_s = 1{,}000$ |
|---|---|---|---|---|
| $n_r = 3$ | .048 | .050 | .045 | .055 |
| $n_r = 5$ | .048 | .053 | .054 | .044 |
| $n_r = 10$ | .054 | .048 | .042 | .050 |
| $n_r = 20$ | .056 | .045 | .050 | .046 |
| Average | .052 | .049 | .048 | .049 |

Table 3
*Power of Tukey's Test: Successful Detection of Nonadditivity in One-Facet Design (1,000 Replications per Condition)*

| | $n_s = 25$ | $n_s = 50$ | $n_s = 100$ | $n_s = 1{,}000$ |
|---|---|---|---|---|
| $n_r = 3$ | .55 | .69 | .75 | .79 |
| $n_r = 5$ | .84 | .91 | .94 | .93 |
| $n_r = 10$ | .97 | .99 | .99 | 1.00 |
| $n_r = 20$ | 1.00 | 1.00 | 1.00 | 1.00 |
| Average | .84 | .90 | .92 | .93 |

measurement. For each $n_s$, results show that the average power of Tukey's test is above .80 across the different numbers of raters, indicating that the test is sensitive to the type of nonadditive interaction suggested by Myers (1979) when it in fact exists in data.

## Empirical Study

### Data

We collected the empirical data in 2009 during a judgmental study of educational standards in Oklahoma (Cook, Wilmes, Chi, & Lin, 2009). This empirical study was motivated by the mandate, under the No Child Left Behind Act (2002), to conduct content alignment studies for English language learners (ELLs) in the U.S. K-12 setting (see Lin & Zhang, 2013, for more detail on ELL content alignment). One of the objectives of the study was to rate the cognitive complexity of a set of 25 English language performance standards using the Depth of Knowledge (DOK) scale developed by Webb (2002). On a scale of 1 to 4, trained raters, who were either content-area teachers or ELL specialists, gave DOK ratings based on the content/task represented in the performance standards. Level 1 is the lowest level, representing low cognitive-demand processing, while level 4 indicates high-level complex processing. A panel of four raters ( $n_r = 4$) rated each performance standard independently based on the established cognitive scale. The standards and raters were crossed in the data set; that is, every rater rated the same set of 25 standards, and the performance standards were the objects of measurement ( $n_s = 25$). In the empirical study, the extent to which the panel of raters reliably interpreted the performance standards in a consistent fashion, with respect to the cognitive scale, was of primary importance because raters' reliability needs to be examined prior to making valid interpretations of the rater-mediated results from the content alignment study. The phi-coefficient, a reliability-like coefficient in G theory, was adopted to serve this purpose and is computed as follows:

$$\text{phi} - \text{coefficient } (\Phi) = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_{sr,e}^2}{n_r}}. \tag{3}$$

Recall from Table 1 that the additive and nonadditive models differ in the estimation of variance component for the objects of measurement ($\hat{\sigma}_s^2$). The underestimation

of $\sigma_s^2$, due to failing to consider nonadditivity when the data is in fact nonadditive, will also result in the underestimation of the phi-coefficient. Following the suggestion by Scheffe (1999) regarding data scrutiny, we first examined the magnitude of the variance component for errors ($\hat{\sigma}_{sr,e}^2$) in the empirical analysis. Upon finding a relatively large proportion for the error component, we performed Tukey's test for nonadditivity. A significant $F_{\text{Tukey}}$ would suggest the presence of nonadditivity in the data and a potential underestimation of $\sigma_s^2$.

## Results

For illustrative purposes, in addition to examining the relative magnitude of $\hat{\sigma}_{sr,e}^2$, we estimated the variance components for standards and raters by using the one-facet additive model regardless of the nature of the data, be it additive or nonadditive. The mean squares, estimated variance components, and their respective proportions of total variance are reported in Table 4.

First of all, it is obvious that the relative magnitude of $\hat{\sigma}_{sr,e}^2$ is large in that it accounts for 47.4% of total variance, which is a hint of potential nonadditivity in the data. Second, it is odd to observe a negative $\hat{\sigma}_s^2$ because it is against the notion of a variance component. It is possible that the negative value is a result of the underestimation of $\sigma_s^2$ in the presence of nonadditivity based on the additive model, which further warrants the use of Tukey's test to detect nonadditivity.

Tukey's single-degree-freedom test for nonadditivity shows a significant nonadditive interaction, $F_{\text{Tukey}}(1,71) = 221.098$, $p < .001$, indicating that the one-facet nonadditive model should have been used instead in the analysis. Table 5 presents

Table 4
*Mean Squares, Estimated Variance Components, and Proportions of Total Variance Based on One-Facet Additive Model*

| | Observed Mean Square | Estimated Variance Component | Proportion of Total Variance |
|---|---|---|---|
| Standard (*s*) | .0892 | −.005 | 0% |
| Rater (*r*) | 3.13 | .121 | 52.6% |
| Error (*sr, e*) | .1092 | .109 | 47.4% |

*Note.* The negative variance component was set to zero in the calculation of proportions.

Table 5
*Mean Squares, Estimated Variance Components, and Proportions of Total Variance Based on One-Facet Nonadditive Model*

| | Observed Mean Square | Estimated Variance Component | Proportion of Total Variance |
|---|---|---|---|
| Standard (*s*) | .0892 | .017 | 6.9% |
| Reviewer (*r*) | 3.13 | .121 | 49.0% |
| Error (*sr, e*) | .1092 | .109 | 44.1% |

the mean squares, estimated variance components, and their respective proportions of total variance based on the one-facet nonadditive model.

Because of the significantly large nonadditive interaction contrast identified by Tukey's test, we observed the underestimation of variance component for the objects of measurement ($\sigma_s^2$) based on the additive model in Table 4 ( $\hat{\sigma}_s^2 = -.005$). We then corrected the underestimation upward based on the nonadditive model presented earlier in the Method section and showed the results ( $\hat{\sigma}_s^2 = .017$) in Table 5. Next, by plugging in the estimated variance components into Equation 3, we obtained the phi-coefficient to assess the reliability of the panel of raters in interpreting the performance standards. Had the additive model been used in the analysis, the phi-coefficient would have been –.095. With the correction of $\hat{\sigma}_s^2$ based on the nonadditive model, the phi-coefficient is .231.

## Discussion

The current study seeks to advance the discussion of nonadditivity in the context of G-theory applications. It has been shown empirically that when nonadditivity in data is present, the variance component for the objects of measurement can be underestimated, and this is one of the possible reasons that leads to negative estimated variance components in practice. More importantly, the current study evaluates the usefulness of Tukey's test under the G-theory framework in detecting nonadditivity in terms of Type I and Type II error rates in a one-facet model, and it further demonstrates the correction for the underestimation of the variance component for the objects of measurement based on Tukey's *F* ratio statistic. In the presence of nonadditivity in data, variance components for the rater effect and for the error term are assumed to be the same between the additive and nonadditive one-facet models in G theory, while the difference between the additive and nonadditive models lies in the variance component for the objects of measurement.

In rater-mediated measurement under the G-theory framework, raters are assumed to be randomly sampled from the *universe of admissible raters*. All raters in this *universe* are well-calibrated and unbiased, in the sense that the raters are interchangeable with one another. Nevertheless, even in a rigorous rater-training system, certain practical realities might reduce its effectiveness, such as unexpected time pressure due to an unforeseen short turnaround time for rating. Some raters' judgments might be more susceptible to time pressure. As a result, although all the raters have been trained, some of them may not be interchangeable with those from the *universe of admissible raters*. In practice, any trained raters are assumed to be in the *universe of admissible raters*; however, this assumption cannot be taken as a given. The current article presents one instance of the violation of the rater interchangeability assumption. Using Tukey's test, we demonstrated its usefulness in evaluating the above assumption in relation to nonadditivity. Most importantly, Tukey's test can serve as a practical screening tool for data scrutiny in rater-mediated measurement. As such, the presence of nonadditivity identified by Tukey's test indicates that some or all of the raters are not interchangeable with those from the *universe of admissible raters*, suggesting that recalibrating the raters and thus re-collecting the data may be warranted. If re-collecting data is not possible, the method presented in this article represents

one way of correcting biased estimation of variance component for the objects of measurement.

The current article is limited to the discussion of nonadditivity in one-facet model in G theory. Future research can broaden the scope by applying and evaluating Tukey's test in two-facet models because the most common facets in many rater-mediated measurement (e.g., constructed-response tests) are those of raters and tasks. Although Tukey's test is not perfect in the sense that "[it] will not be sensitive to all interactions" (Myers, 1979), it is nevertheless an effort to address nonadditivity given the complications introduced by it. Future research can aim to investigate the test's sensitivity (or lack thereof) to various types of nonadditive interactions and to develop other procedures that can complement Tukey's test when it fails to detect nonadditivity.

## References

Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, *5*, 141–160.

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

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

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Applied Psychological Measurement*, *55*, 157–176.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, *4*, 219–240.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2008). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, *22*, 1–21.

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2014). An examination of the replicability of Angoff standard setting results within a generalizability theory framework. *Journal of Educational Measurement*, *51*, 127–140.

Cook, G., Wilmes, C., Chi, Y., & Lin, C. (2009). *Alignment between the Oklahoma Priority Academic Student Skills and the WIDA Consortium English Language Proficiency Standards*. Madison: Wisconsin Center for Education Research, University of Wisconsin at Madison.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. NewYork, NY: Wiley.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York, NY: Springer.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, *7*, 323–342.

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

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Learning.

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly*, *7*, 219–233.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

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

Lin, C.-K., & Zhang, J. (2013). Enhancing standard-based validity for ELL population: A perspective from correspondence between standards. *TESOL Quarterly*, *47*, 399–410.

Lin, C.-K., & Zhang, J. (2014). Investigating correspondence between language proficiency standards and academic content standards: A generalizability theory study. *Language Testing*, *31*, 413–431.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston, MA: Allyn and Bacon.

No Child Left Behind Act of 2001. (2002). Pub. L. No. 107–110, 3 U.S.C. (2002).

Scheffe, H. (1999). *The analysis of variance*. New York, NY: Wiley.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232.

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

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239–261.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, *5*, 232–242.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

Yin, P., & Sconing, J. (2008). Estimating standard errors of cut scores for item rating and mapmark procedures: A generalizability theory approach. *Educational and Psychological Measurement*, *68*, 25–41.

Zhang, J., & Lin, C.-K. (2016). Generalizability theory with one-facet nonadditive models. *Applied Psychological Measurement*, *40*, 367–386.

## Authors

CHIH-KAI (CARY) LIN is a Psychometrician at American Institutes for Research, 1000 Thomas Jefferson St. NW, Washington, DC 20007; clin@air.org. His primary research interests include generalizability theory, item response theory, value-added modeling, and standard setting.

JINMING ZHANG is an Associate Professor at the University of Illinois at Urbana-Champaign, 236A Education Building, 1310 South 6th St., Champaign, IL 61820; jmzhang@illinois.edu. His primary research interests include multidimensional item response theory, generalizability theory, dimensionality assessment techniques, large-scale assessments, computerized adaptive testing, cognitive diagnostic modeling, and test security.