A Multivariate Generalizability Theory Approach to Standard Setting

Applied Psychological Measurement

2015, Vol. 39(7) 507–524

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615577972

apm.sagepub.com

Yi-Fang Wu¹ and Hueying Tzou²

Abstract

Generalizability theory (G theory) allows researchers to assess the many sources of variance inherent in complex standard setting procedures involving the determination of cut scores. The flexibility of G and D studies provides a way to conceptualize and quantify the results of different standard settings once the universe of admissible observations and the universe of generalization are defined. The current article applies a multivariate single-facet design for estimating standard errors of cut scores. For practical purposes, several multivariate D study designs are used to investigate what effect various panel sizes and test lengths have on the precision of the standard setting process. The current study demonstrates the advantages and usefulness of multivariate G theory in determining the accuracy of cut scores in practical applications of standard setting procedures.

Keywords

multivariate generalizability theory, standard setting, the modified Angoff method, standard errors

Introduction

Setting performance standards in elementary and secondary education has been an integral part of the development of statewide assessments mandated by the No Child Left Behind Act (NCLB; 2002). In addition, setting cut scores is essential in the development of licensure and certificate examinations. Earlier literature viewed standard setting as an outcome variable of the test development process, but a more recent, contemporary perspective (e.g., Cizek, 2012a, 2012b; Cizek & Bunch, 2007) advocates treating the standard setting method as a design parameter. The goals of the standard setting process, the standard setting method chosen, as well as the characteristics of the standard setting participants should be tightly aligned with the purpose and design of the test.

Despite the political, social, and economic complexity of any specific application of the standard setting process, it is important to gather evidence for evaluating the precision of the results

¹University of Iowa, Iowa City, USA

²National University of Tainan, Taiwan

Corresponding Author:

Yi-Fang Wu, Educational Measurement and Statistics, University of Iowa, 361 Lindquist Center, Iowa City, IA 52242, USA.

Email: yi-fang-wu@uiowa.edu, wuyifang91@gmail.com

of the standard setting procedure. To evaluate standard setting results (i.e., cut scores), it is critical to report both validity and reliability evidence for assuring the adequacy and appropriateness of the performance standards. Hambleton and Pitoniak (2006) indicated that the internal constituent of the standard setting evaluation elements emphasizes the consistency within method by inspecting the precision of the estimate of the cut scores. As will be demonstrated, multivariate generalizability theory (G theory) can provide such quantitative evidence.

Early literature by Brennan (1995), Brennan and Lockwood (1980), and Kane and Wilson (1984) clarified the application of G theory to standard setting. As outlined by Brennan (1995), the key question in evaluating standard setting results is “how variable would the cut scores be if the process were replicated?” (p. 271). To answer this question, later researchers focused on the standard errors (*SEs*) of cut scores from different standard setting procedures. For example, in some medical examinations, researchers applied G and D study designs, which considered rater and item facets in the Angoff-based procedures (e.g., Verhoeven et al., 1999; Verhoeven, Verwijnen, Muijtens, Scherpbier, & van der Vleuten, 2002) and in the borderline regression method (e.g., Kramer et al., 2003). Others discussed more complex designs that went beyond rater and item facets (e.g., Chang, 1999; Chang & Hocevar, 2000).

Because the Angoff-based methods are classic standard setting procedures and are often implemented for large-scale statewide assessments as well as licensure and certificate examinations, researchers have focused on using G theory for estimating the *SEs* of cut scores (Arce-Ferrer & Yin, 2007; Yin & Sconing, 2006, 2008), the stability of cut scores across rounds (Clauser et al., 2009), or both (Tzou, Wu, & Lin, 2008; Wu & Tzou, 2010). The popularity of item response theory (IRT) has led to the development of other standard setting methods, such as the bookmark method. Lee and Lewis (2001, 2008) have investigated the application of G theory for estimating the *SEs* of cut scores in the bookmark standard setting.

Although past studies used G theory in the standard setting analyses, none have considered multivariate G theory. Statewide or nationwide assessments contain different content categories, usually with a different set of items within each content category. In licensure and certificate examinations, tasks diverse and the number of tasks associated may vary by skills. Taking into account these multivariate aspects of such assessments leads naturally to analysis by multivariate G theory. For each object of measurement, there are multiple universe scores, each of which is associated with a level of one or more fixed contents, and the items nested with each fixed levels are randomly parallel. That is, the design is mixed with the random item factor nested within the fixed content category factor. Using multivariate G theory can appropriately identify all sources of error and accurately evaluate their contributions to the *SEs* of cut scores. In addition, it is possible that each category or skill has unequal number of items or tasks, which creates an unbalanced design and increases the complexity of the analyses. Multivariate G study designs deal with both balanced and unbalanced situations and they are much simpler than univariate G study designs for unbalanced cases (e.g., Brennan, 2001a), and the computer program mGENOVA (Brennan, 2001b) can be used to analyze a variety of multivariate G study designs.

Motivated by the above considerations, the purpose of the current study was to demonstrate a method of quantifying *SEs* of cut scores when standard setting data involve fixed and random effects. A multivariate G theory approach was applied to estimate the *SEs* of the cut scores resulted from the modified Angoff standard setting method. In the authors' definition, the universe of admissible observations (UAOs) and the universe of generalization (UG) had the same structure. A single-facet multivariate G study design was demonstrated to estimate variance-covariance components after which several D studies with various numbers of panelists and items were analyzed to understand how varied the cut scores could be if the modified Angoff procedures were replicated under different but randomly parallel conditions. The *SEs* of the cut

Table 1. Data Structures for Implementing the Modified Angoff Standard Setting Procedures.

Content	Algebra	Statistics	Geometry	Measurement	Number sense and computation	Overall
No. of items scored ^a	11	10	10	31	42	104
Average <i>p</i> values ^b	.452	.600	.468	.536	.515	.518
No. of items rated	11	10	10	30	40	101
Average <i>p</i> values ^c	.452	.600	.468	.542	.526	.524
No. of panelists	12	12	12	12	12	

^aThe number of items scored is also the operational number of item.

^bThese conventional *p* values are calculated from the number of items scored by content.

^cThese conventional *p* values are calculated from the number of items rated by content.

scores were then considered. The formulas for the *SEs* were derived and their applicability was demonstrated to the modified Angoff standard setting results.

Angoff-based methods require the panelists to provide item-level performance estimates for borderline groups, and this presents the panelists with quite a lot of work. To reduce this workload, an item-grouping approach was incorporated by grouping items in cells of similar difficulty and examined its impact on the accuracy of the cut scores. The D studies were also used to investigate whether a reduction in the number of items rated by each panelist would affect measurement precision.

In an effort to investigate the accuracy and stability of the cut scores as well as to illustrate the practicability of applying a multivariate G theory approach to standard setting, the current study used empirical standard setting data to answer the following questions:

Research Question 1 (RQ1): What was the variability of the basic, proficient, and advanced cut scores?

Research Question 2 (RQ2): How did the *SEs* of the cut scores change in different standard setting rounds?

Research Question 3 (RQ3): What impact did panelist size and test length have on the *SEs* of the cut scores?

It is expected that the application of multivariate G theory will provide practical information for interpreting results and for planning future standard settings.

Method

Standard Setting Process

Materials. The data were from a standard setting study for determining the cut scores for a national fourth-grade mathematics achievement test. Students' performances were classified into four levels (i.e., below basic, basic, proficient, and advanced) by three cut scores. The test consisted of 104 operational multiple-choice items within five content categories: algebra, statistics, geometry, measurement, and number sense and computation. As in Table 1, the average item difficulties (i.e., conventional *p* values) of the five categories are .452, .600, .468, .536, and .515 based on 11, 10, 10, 31, and 42 items, respectively. The overall test difficulty is .518 based on a sample size of 8,120 students. When the math test was scored by a three-parameter logistic IRT (3PL IRT) model, three items had unusual characteristics (e.g., extreme *b*

Figure 1. The multistage standard setting procedure.

parameter estimates) and then were excluded from both the item pool and the standard setting. Thus, the numbers of items within the five content categories used for setting the cut scores were 11, 10, 10, 30, and 40, respectively.

Panelists. Twelve panelists participated in the standard setting. One was a new assistant professor in mathematics education and the others were elementary school math teachers. Their teaching experience in math was 8 years on average. The assistant professor who had been an elementary school teacher for more than 10 years did not behave differently from the other panelists. In addition, most had prior experience in the modified Angoff procedures.

Procedure. The agenda of the modified Angoff standard setting procedures was guided by Angoff (1971), the National Assessment of Educational Progress (Allen, Jenkins, Kulick, & Zelenak, 1997), Reckase (1998, 2000), Hambleton (2001), Hambleton and Pitoniak (2006), and Cizek and Bunch (2007). The panel participated in the standard setting which consists of three rounds in three consecutive days. All items were rated by each panelist; that was, each panelist provided item-level performance estimates for the borderline groups of basic, proficient, and advanced levels. To investigate item-grouping effect (i.e., prior to being rated, items had been grouped with respect to their item difficulties), six panelists were randomly selected and then exposed to the effect. These six panelists who were unaware of the effect formed a group called the “item-preordered group” (IPG) whereas the other six panelists who were not introduced the effect constituted the “non-item-preordered group” (NIPG). Except for the item grouping, the two groups followed exactly the same standard setting procedures.

Figure 1 shows the modified Angoff standard setting procedures integrated with regular “normative and reality feedback” (Cizek & Bunch, 2007, pp. 54-56) as well as other strategies for improving interpanelist and intrapanelist consistency. To improve the interpanelist consistency, the normative feedback was done by providing panelists the information about one’s ratings in comparison with other panelists’ ratings. The means and standard deviations of the panelists’ ratings on the score scale (i.e., the estimated minimal passing scores) and a correlation matrix that contained the pairwise correlations among panelists’ ratings were provided. The reality feedback offered the empirical item difficulty (i.e., conventional p value)

such that panelists could figure how their judgments on the items were compared with actual examinee performances. To improve the intrapanelist consistency at the final round, every panelist was given a customized chart where his or her panelist's ratings at the second round were listed in the order of empirical item difficulties. Panelists then reviewed the items with similar p values and checked their ratings to detect discrepancy, and finally made adjustments if necessary. In short, different types of feedback were intended to decrease the inconsistency within the panel and guarantee the validity of the final cut scores (Hambleton & Pitoniak, 2006).

A Multivariate G Theory Approach

A multivariate G theory approach was applied to analyze panelists' item ratings from the standard setting. Under the multivariate G theory framework, the SE of the final average rating ($\bar{X}$) was explored, which could be conceptualized as the SE of the cut score on the rating scale (e.g., the p value scale). The associated facets, a multivariate G study $p^{\bullet} \times i^{\circ}$ design, and the D study designs are discussed below.

Facet. Cut scores are determined after the final standard setting round. Thus, factors involved in the standard setting procedure are possible facets in a G theory model. First, the *item facet* contributes to error; it is a random facet because a different set of items can be involved in each replication. Second, *panelist facet* is essential in assessing the adequacy of certain standard setting procedures like the Angoff-based method (Brennan, 1995). It contributes to the SE s of the cut scores; the panelist effect is also random. Third, the math test is composed of five content categories, and thus, the *content (category) facet* that is fixed is considered. The content category facet is fixed because every replication of the standard setting procedure involves the same categories. Finally, it is noted that round is not regarded as a facet. In standard setting, rounds are not randomly parallel, especially when the standard setting incorporates different feedback procedures. In short, the multiple standard setting rounds are fixed procedures and they are not interchangeable. As Brennan (1995) emphasized, the entire standard setting procedure should be viewed as one replication as the final cut scores depend on the final round only. The variability over items and the panelists are of the most interest.

Multivariate G study design. To clearly identify SE s of measurement, the specification of the UAO is now defined. As stated, the content category facet is fixed. The items are random in the sense that each replication of the standard setting procedure could involve a different set of items within each level of the content category facet. Every item is rated by each panelist. In a univariate G study sense, this is a $p \times (i: h)$ design, where p , i , and h represent panelist, item, and content facets, respectively. Note that the unequal numbers of items within each content category make the standard setting data unbalanced. The fixed content facet and the random-effects variance components design associated with each fixed content level yields a multivariate G study design, $p^{\bullet} \times i^{\circ}$, with the number of levels for the fixed facet being n_h . The solid circle, $\bullet$, designates that the panelist facet is crossed with the fixed multivariate variable (i. e., content) whereas the empty circle, $\circ$, designates that the item facet is nested within the fixed multivariate variable. In other words, there is a random-effects $p^{\bullet} \times i^{\circ}$ design within each of the five fixed content categories and also, any single item is only associated with one single content category.

The standard setting data have five content categories (i.e., $n_h = 5$), and the model equations for different categories of items, $h_{(1)}$, $h_{(2)}$, $h_{(3)}$, $h_{(4)}$, and $h_{(5)}$, are presented as follows:

Table 2. Numbers of Items by Content for the D Studies.

Content	Test length proportion								
	Original	50%	60%	70%	80%	90%	140%	160%	200%
Algebra	11	5	6	7	8	9	14	16	20
Statistics	10	5	6	7	8	9	14	16	20
Geometry	10	5	6	7	8	9	14	16	20
Measurement	30	15	18	21	24	27	42	48	60
Number sense and computation	40	20	24	28	32	36	56	64	80
Total	101	50	60	70	80	90	140	160	200

$$\begin{aligned} X_{pih(1)} &= \mu_{h(1)} + \nu_{(1)p} + \nu_{(1)i} + \nu_{(1)pi}, \\ X_{pih(2)} &= \mu_{h(2)} + \nu_{(2)p} + \nu_{(2)i} + \nu_{(2)pi}, \\ X_{pih(3)} &= \mu_{h(3)} + \nu_{(3)p} + \nu_{(3)i} + \nu_{(3)pi}, \\ X_{pih(4)} &= \mu_{h(4)} + \nu_{(4)p} + \nu_{(4)i} + \nu_{(4)pi}, \\ X_{pih(5)} &= \mu_{h(5)} + \nu_{(5)p} + \nu_{(5)i} + \nu_{(5)pi}. \end{aligned} \tag{1}$$

In Equation 1, $\nu_{(1)}$, designates effects for the first content category, $h_{(1)}$; $\nu_{(2)}$ designates effects for $h_{(2)}$; and so forth. It follows that the variance and covariance components for the population and the UAO can be grouped into three symmetric matrices: $\sum_p$, $\sum_i$, and $\sum_{pi}$. As all panelists contribute data to all levels of h but items are nested in different levels of h , it turns out that $\sum_p$ is a full matrix but $\sum_i$ and $\sum_{pi}$ are diagonal.

Multivariate D study designs. To understand the impact of panel size and test length on the *SEs* of the cut scores, several multivariate D studies were conducted, considering different numbers of panelists and items by varying n'_p and n'_{ih} . The D study designs included 15 panel sizes (i.e., $n'_p = 2, 4, 6, \dots$, and 30) and nine test lengths, including the original test length, half and doubled test lengths, and increasing proportions of the original test length (i.e., 60%, 70%, 80%, 90%, 140%, and 160%). The test length was varied proportionally with respect to content categories because the relative importance of each content category should remain the same as the test specification. The numbers of items by content category in the multivariate D study $p^\bullet \times i^\circ$ designs are displayed in Table 2.

The SE of the cut score under the $p^\bullet \times i^\circ$ design. The *SE* of the cut score under the $p^\bullet \times i^\circ$ design, denoted $SE(\bar{X})$, is the *SE* involved in using the mean over the sample of both panelists and items as an estimate of the mean over both the population of panelists and the universe of items. In the following, the derivation of $SE(\bar{X})$ under the $p^\bullet \times i^\circ$ design is presented with the number of the fixed content facet being n'_{ih} and $n'_h = 5$.

The $SE(\bar{X})$ is the *SE* of measurement for mean ratings over all n'_p panelists and $n'_{i\bullet}$ items, where $n'_{i\bullet} = n'_{ih(1)} + n'_{ih(2)} + n'_{ih(3)} + n'_{ih(4)} + n'_{ih(5)}$, and $n'_{ih(1)}$ represents the number of items within the first content category (i.e., algebra) and so on. Followed by the error variance for the mean in a univariate $p \times I$ design and by the definition of the variance for the composite (e.g., Brennan, 2001a), the $SE(\bar{X})$ over both items and panelists is defined in the following way.

First, let $\sum_p = \sum_p / n'_p$. Also, define $\sum_I = \sum_i / \sum_{n_{ih}}$; that is, divide the h th diagonal element in $\sum_i$ by the corresponding number of items, n_{ih} . Then, define $\sum_{pI} = \sum_{pi} / n'_p \sum_{n_{ih}}$ by dividing

the h th diagonal element in $\sum_{pi}$ by $n'_p n'_{ih}$. Note that $\sum_p$ is the variance–covariance component matrix among panelists, $\sum_i$ is the variance–covariance component matrix for items within content category, and $\sum_{pi}$ is for the item–panelist interaction within content category. Intentionally, $\sum_{n_{ih}}$ is designed to be a diagonal matrix containing the numbers of items within the levels of h (e.g., $\sum_{n_{ih}} = \text{diag}(11, 10, 10, 30, 40)$ in this study). In addition, nominal weight (w_h) (e.g., Wang & Stanley, 1970) reflects the proportion of the number of items in the measurement procedure that is associated with each h to the total test. It is commonly defined as $w_h = n_{ih}/n_{i\bullet}$, where $n_{i\bullet}$ designates the total number of items over all categories of the content facet. The nominal weights are applied to the corresponding variances (σ_h^2) or covariances ($\sigma_{hh'}$) to get the SE ,

$$SE(\bar{X}) = \left\{ \left[\sum_{h=1}^{n_h} w_h^2 \sigma_h^2(P) + \sum_h \sum_{h' \neq h}^{n_h} w_h w_{h'} \sigma_{hh'}(P) \right] + \sum_{h=1}^{n_h} w_h^2 \sigma_h^2(I) + \sum_{h=1}^{n_h} w_h^2 \sigma_h^2(PI) \right\}^{\frac{1}{2}}. \quad (2)$$

In Equation 2, the sum of the two pieces in the square bracket represents the weighted sum of all elements in the $\sum_p$ while the third and the last pieces represent the weighted sums of the diagonal elements of $\sum_i$ and $\sum_{pi}$, respectively. The estimated variance–covariance component matrices can be denoted by $\widehat{\sum}_p$, $\widehat{\sum}_i$, and $\widehat{\sum}_{pi}$ and be obtained using mGENOVA (Brennan, 2001b); an example code is presented in the Appendix. Followed by the definitions above and Equation 2, one can get the estimates of the SE s for the Angoff cut scores. The development of Equation 2 shows how different sources of variances contribute to the SE s of cut scores in an explicit way. A generic form of the SE s can also be referred to (see Equation 4.20 in Brennan, 2001a). It is for obtaining the error variance associated with using the observed grand mean of ratings, $\bar{X}$, as an estimate of the mean in the population and universe. The estimated SE s can also be provided by mGENOVA using the verbal identifier “Composite Error Standard Deviation of Mean.”

Results and Discussion

G Study Results—Quantifying the Variability of the Cut Scores

The panel was divided into two groups, IPG and NIPG, for investigating the item-grouping effect. The performance standards provided by one group could not be generalized to those by the other because IPG and NIPG was not randomly parallel to one another. As this group effect was fixed, two multivariate single-facet G study designs have been applied to IPG and NIPG, respectively.¹ The item-grouping effect could be examined by comparing the G study results from IPG to those from NIPG.

Table 3 presents the estimated variance–covariance matrices of panelist ($\widehat{\sum}_p$), item ($\widehat{\sum}_i$) as well as interaction ($\widehat{\sum}_{pi}$) facets for the basic, proficient, and advanced cut scores; the matrices were estimated from Round 3 standard setting results. As shown, for IPG, within each content facet, the diagonal elements of $\widehat{\sum}_p$ were smaller than the diagonal elements of $\widehat{\sum}_i$, indicating that the variability among panelists was less than the variability among items. The item–panelist interactions were quite small and they were smaller than they were from Rounds 1 and 2, indicating that the feedbacks provided in the standard setting procedure improved the panel consistency.

Similar findings were found for NIPG. Comparing the estimated matrices between IPG and NIPG, the differences suggested that the item-grouping effect do exists. For item effect, the variance estimates from IPG were uniformly smaller than the variance estimates from NIPG irrespective of content areas. In other words, the item variability from IPG was less than the item variability from NIPG. This was found for each cut score, implying that the item-grouping effect was functioning when the panelists of IPG provided their item ratings at each

Table 3. The Estimated Variance–Covariance Component Matrices for Basic (B), Proficient (P), and Advanced (A) Cut Scores From the Multivariate G Study $p \times i$ Design for IPG and NIPG.

IPG			
B			
P			
A			
$\hat{\Sigma}_p =$	$\begin{bmatrix} .0001 & .0005 & .0000 & .0005 & .0000 & .0001 \\ .0003 & .0000 & .0005 & .0000 & .0001 & .0000 \\ .0000 & .0004 & .0000 & .0000 & .0001 & .0000 \\ .0001 & .0006 & .0000 & .0003 & .0004 & .0000 \\ .0140 & .0154 & .0263 & .0187 & .0132 & .0029 \end{bmatrix}$	$\begin{bmatrix} .0076 & .0039 & .0060 & .0085 & .0064 & .0086 & .0061 & .0045 & .0064 & .0044 & .0066 & .0048 & .0068 & .0049 & .0051 \end{bmatrix}$	$\begin{bmatrix} .0075 & .0037 & .0056 & .0075 & .0052 & .0067 & .0056 & .0039 & .0052 & .0039 & .0063 & .0044 & .0059 & .0045 & .0050 \end{bmatrix}$
$\hat{\Sigma}_i =$	$\begin{bmatrix} .0124 & .0300 & .0189 & .0159 \end{bmatrix}$	$\begin{bmatrix} .0122 & .0042 & .0200 & .0139 & .0093 \end{bmatrix}$	
$\hat{\Sigma}_{pi} =$	$\begin{bmatrix} .0063 & .0060 & .0050 & .0070 & .0052 \end{bmatrix}$	$\begin{bmatrix} .0041 & .0041 & .0046 & .0041 & .0032 \end{bmatrix}$	
NIPG			
B			
P			
A			
$\hat{\Sigma}_p =$	$\begin{bmatrix} .0015 & .0039 & .0015 & .0022 & .0013 & .0023 & .0034 & .0021 & .0033 & .0022 & .0033 & .0020 & .0031 & .0028 \end{bmatrix}$	$\begin{bmatrix} .0023 & .0012 & .0019 & .0010 & .0013 & .0020 & .0010 & .0016 & .0016 & .0016 & .0020 & .0011 & .0016 & .0017 & .0018 \end{bmatrix}$	$\begin{bmatrix} .0039 & .0033 & .0037 & .0029 & .0029 & .0036 & .0029 & .0032 & .0030 & .0032 & .0040 & .0032 & .0035 & .0034 & .0038 \end{bmatrix}$
$\hat{\Sigma}_i =$	$\begin{bmatrix} .0240 & .0345 & .0252 & .0175 \end{bmatrix}$	$\begin{bmatrix} .0219 & .0368 & .0295 & .0188 \end{bmatrix}$	$\begin{bmatrix} .0147 & .0286 & .0235 & .0137 \end{bmatrix}$
$\hat{\Sigma}_{pi} =$	$\begin{bmatrix} .0024 & .0013 & .0032 & .0022 & .0017 \end{bmatrix}$	$\begin{bmatrix} .0030 & .0019 & .0025 & .0027 & .0027 \end{bmatrix}$	$\begin{bmatrix} .0048 & .0035 & .0033 & .0050 & .0047 \end{bmatrix}$

Note. Lower diagonal elements in the panelist matrices are covariances. IPG = item-preordered group; NIPG = non-item-preordered group.

Figure 2. The SEs of basic, proficient, and advanced cut scores (on the metric of item proportion correct) at three standard setting rounds under the multivariate $p \times i^\circ$ design for IPG and NIPG.
 Note. IPG = item-preordered group; NIPG = non-item-preordered group.

achievement level. The item–panelist interactions were quite similar in both groups, except for the proficient level. As to the panelist matrices, the variance–covariance estimates from IPG were larger than the variance–covariance estimates from NIPG for proficient and advanced cut scores. The panelist effect seemed to be stronger in IPG, which implied that there was a lack of consensus among the IPG panelists with respect to their opinion about the borderline groups at the proficient and advanced levels. On the contrary, the NIPG panelists have reached a greater agreement in their opinion on the two borderline groups. The panelist effect shown for the basic cut score told a very different story. The reason why the variance–covariance estimates of the panelist matrix from IPG were uniformly smaller than those from NIPG at the basic level needs a further investigation.

In sum, the variability among items dominated the variability of the cut scores because the estimated variances of the item facet were larger than the variances of the panelist facet and the interactions. Also, the effect of grouping items in cells of similar difficulty was reflected by less variability among item ratings from IPG than NIPG.

Change of the SEs Across Different Standard Setting Rounds

As defined previously, the UAO and the UG had the same structure. Thus, the SEs were estimated from the same number of panelists and items in the G study designs (i.e., $n'_p = n_p = 6$, $n'_{ih(1)} = n_{ih(1)} = 11$, $n'_{ih(2)} = n_{ih(2)} = 10$, $n'_{ih(3)} = n_{ih(3)} = 10$, $n'_{ih(4)} = n_{ih(4)} = 30$, and $n'_{ih(5)} = n_{ih(5)} = 40$). Using the $\widehat{\Sigma}_p$, $\widehat{\Sigma}_i$, and $\widehat{\Sigma}_{pi}$ from Round 3, the SEs of the basic, proficient, and advanced cut scores from IPG and NIPG were computed, respectively. The change of the SEs from various rounds could tell whether the feedbacks and the modifications had worked for the modified Angoff method. Figure 2 shows the SEs of the basic, proficient, and advanced cut scores at each round under the multivariate $p \times i^\circ$ design.

First, the *SE* of the basic cut score from IPG in Round 1 was .0401—much higher than the rest *SEs*. In the standard setting, panelists were asked to provide the item-by-item Angoff estimates for the borderline examinees of the basic level first, followed by the proficient and advanced levels. For IPG, setting the Round 1 cut score for the basic level was the first time they examined all the items. Rating so many items might cause the panelists mentally burdened, which introduced instability to the panel. In the panelists' discussion followed by Round 1, some panelists reported heavy loads in the beginning and found themselves accommodated, especially when they provided the Angoff estimates for the proficient and advanced borderline groups. Although not reported, the $\widehat{\Sigma}_p$ of the basic level in Round 1 for IPG contained the largest variance–covariance estimates among all panelist matrices, indicating that the panelists of IPG varied in their averaged ratings for the basic borderline group. As the variances and covariances in $\widehat{\Sigma}_p$ went into the *SE* as Equation 2 shows, it was not surprising that the *SE* of the basic cut score in Round 1 was the highest. Interestingly, the findings above did not apply to NIPG. In Round 1, the *SE* of the basic cut score (i.e., .0201) was not much different from the *SE* of the proficient cut score (i.e., .0193) or the advanced cut score (i.e., .0204).

The *SEs* of the cut scores from Rounds 1 and 2 fluctuated at different achievement levels but finally, for most cases, they minimized in Round 3. The exceptions included the advanced cut score from IPG, where a suspiciously low *SE* = .0130 was obtained in Round 1, and the basic cut score from NIPG, where a slightly low *SE* = .0169 was obtained in Round 2. In general, the panelists' discussion and the normative and reality feedbacks that were used to improve the panel consistency worked as expected. In Round 3, smaller *SEs* were reported for the basic and advanced cut scores (i.e., .0132 and .0156) than the proficient cut score (i.e., .0179) from IPG and roughly the same size of *SEs* were reported from NIPG (i.e., .0172, .0170, and .0168 for basic, proficient, and advanced levels). In sum, provided the higher *SE*, it seemed that the proficient cut score was not as consistent as the basic and advanced cut scores from IPG. However, the three cut scores from NIPG reached almost the same precision.

D Study Results—Decisions on Panel Size and Test Length

In the D studies, the *SEs* of basic, proficient, and advanced cut scores were computed by varying D study sample sizes and using the estimated variance and covariance component matrices from the G study. Only the variance and covariance component matrices from Round 3 were used for reporting the D study results.

As expected, the *SE* became smaller as the panel size enlarged. For IPG, the *SE* lines became flat after $n'_p = 10$ for the basic cut score regardless of test length. For the proficient and advanced cut scores, the *SE* lines became flatter after $n'_p = 22$. In the G study analyses, the $\widehat{\Sigma}_p$ of the basic level had smaller variance–covariance estimates than the $\widehat{\Sigma}_p$ of the other two levels. As there was less panelist variability at the basic level than the proficient and advanced levels, increasing the panel size would have little impact on the deduction of *SEs* at the basic level.

The D study designs also took test length into consideration. To maintain the relative importance of the five content domains, the D study test lengths varied proportionally. For example, the “half” test length had a total of 50 items, where $n'_{ih(1)} = n'_{ih(2)} = n'_{ih(3)} = 5$, $n'_{ih(4)} = 15$, and $n'_{ih(5)} = 20$. It was assumed that there were five randomly parallel items in analysis, statistics, and geometry, respectively, 15 items in measurement, and 20 items in number sense and computation.

As increasing panel size reduced the *SEs*, varying test lengths had an impact on the *SEs* as well. As in Figures 3 and 4, longer tests led to smaller *SEs* than shorter tests as expected. Regardless of panel size, under half of the original test length, the *SEs* were increased by .005 at different achievement levels. However, doubling the test length reduced the *SEs* to a

Figure 3. The SEs of basic, proficient, and advanced cut scores for different D study sample sizes for IPG.

Note. The number represents the test length in proportion to the original test (e.g., “.5” is for the test which is half of the original test length). The numerical string represents the numbers of items within different content categories. IPG = item-preordered group; UG = universe of generalization; UAO = universe of admissible observation.

^aA, S, G, M, and N denote Algebra, Statistics, Geometry, Measurement, and Number sense and computation.

Figure 4. The SEs of basic, proficient, and advanced cut scores for different D study sample sizes for NIPG.

Note. The number represents the test length in proportion to the original test (e.g., “.5” is for the test which is half of the original test length). The numerical string represents the numbers of items within different content categories. NIPG = non-item-preordered group; UG = universe of generalization; UAO = universe of admissible observation.

maximum of .005, especially when the panel size also increased. Although increasing the test length resulted in much more improvement in the measurement precision than enlarging the panel size did, doubling the test length might not be realistic in practice. In addition, the fact that both 80% and 90% *SE* lines were not deviated much from the *SE* line of the original test length indicated that allowing panelists to rate fewer items in standard setting can be a possibility (assuming content representativeness of items is not an issue). Similar results were found for NIPG, except that the *SE* lines became level for all cut scores after $n'_p = 22$ or 24. To be informative, Table 4 summarizes the desirable number of panelists under selected test length proportions, given that the *SE*s of the cut scores were either smaller than .020 or .015.

To reduce *SE*s from .020 to .015 for IPG, one might need to, at least, double the number of panelists. The only exception was for the basic cut score with length proportions of 1.4 and more. Not surprisingly, when the test length proportion was small, a bigger panel was required to result in more consistent cut scores (i.e., $SE < .015$). At the proficient level, in particular, a large panel containing more than 30 panelists was desired if $SE < .015$ and the full test (or fewer items) was rated. But the panel size required for an acceptable *SE* at the basic level was small, due to the fact that the resulting estimated *SE* of the basic cut score from the G study analysis was quite small (i.e., .0132). In practice, however, a standard setting panel containing two experts is unrealistic because a representative sample of panelists usually enlarges the panel size.

For NIPG, reducing *SE*s from .020 to .015 required twice, thrice, or more of the panel size especially for small test length proportions. For the full test or tests with fewer items, a large panel containing more than 30 panelists was needed to achieve $SE < .015$ at all levels. Furthermore, by inspecting the differences between the results of IPG and NIPG, it was found that when $SE < .020$, the desirable numbers of panelists for IPG and NIPG were usually 10 or less; in some cases, NIPG required more panelists under selected test length proportions for a certain levels (e.g., basic level under 80% of the test length) but sometimes it was IPG that required more (e.g., proficient level under 80% of the test length). When $SE < .015$, NIPG tended to need more panelists than IPG except when long tests (e.g., 160% or more) were used at the proficient level. In conclusion, building a larger standard setting panel for NIPG than for IPG was necessary no matter how many items would be rated.

Conclusions and Limitations

The complexity of an empirical standard setting study sometimes makes it hard to quantify the results. The nature of G theory makes itself a powerful tool in terms of detecting different sources of errors though it matters how researchers conceptualize the UAOs (e.g., items). The flexibility of G studies and D studies allows researchers to design their future standard setting agenda.

In this study, a multivariate G study $p \bullet \times i^\circ$ design was used to analyze the standard setting results from the modified Angoff procedures. It was intended to investigate the *SE*s of the cut scores because *SE* is vital when standard setting results are evaluated. As the *SE*s of the basic, proficient, and advanced cut scores for the fourth-grade math assessment were quite small, the final cut scores were acceptable in terms of measurement precision. Most importantly, the fact that the smallest *SE*s from Round 3 guaranteed that the modified Angoff standard setting procedures had worked well. For those cases where the resulting *SE*s from Round 3 were not the smallest, they still provided valuable information. For example, they triggered a key question in evaluating standard setting results (Cizek, 2012a): Did the method, as delivered, deviate in unexpected ways? If that were the case, a close examination of the procedures must be done and certain modifications would be necessary for future standard settings.

Table 4. The Desirable Number of Panelists Under Various Test Length Proportions for IPG and NIPG.

Group level	Test length proportion											
	0.8			1.0			1.4			1.6		
	B	P	A	B	P	A	B	P	A	B	P	A
IPG												
SE < .020	2	10	8	2	8	6	2	6	6	2	6	6
SE < .015	8	>30	20	4	>30	16	2	18	12	2	16	12
NIPG												
SE < .020	8	6	8	6	4	6	6	4	6	4	4	6
SE < .015	>30	>30	>30	>30	>30	>30	24	28	18	12	8	12
										10	6	10

Note. Table values are the number of desired panelists. IPG = item-preordered group; NIPG = non-item-preordered group. B = basic; P = proficient; A = advanced.

The D study results assisted in practical decisions, if the same UG were applicable. The current study found that the desirable number of panelists had better no less than 10. Brandon (2004) also indicated that the desirable panel size of the modified Angoff method was at least 10. Based on the findings, it was also concluded that recruiting more panelists may not result in significant improvement in measurement precision (i.e., $SE < .015$, a decrease of .005 only). Finally, if one were to replicate the modified Angoff standard setting procedures, 80% or 90% of the test length could be considered, because the SE lines of the original test were not discrepant much from the SE lines of the 80% or 90% of the test length. With such an adjustment, the complexity of cognitive burden for panelists could be reduced. In addition, item grouping could lend a hand to improve the precision of the cut scores based on the findings that the SE s from the item-preordered group were smaller than the SE s from the non-item-preordered group. Other things being equal, the desirable panel sizes were smaller for the item-preordered group than the non-item-preordered group under most test length conditions and achievement levels. Nevertheless, for practical considerations, the acceptable magnitude of SE s still needs some deliberation, because they should depend on the real context as well as the G or D study designs. To evaluate, it is also of interest to study the SE s or confidence intervals for the estimated variance components in the future.

In the study, a multivariate G study crossed design was fully discussed and demonstrated. It is also possible to apply a nested design, such as $(p^\bullet : s^\bullet) \times i^\circ$ and $(P^\bullet : S^\bullet) \times I^\circ$ design, when a single panel is split to form smaller independent subpanels. For example, following a modified, extended Angoff method, and a counterbalanced design, Tannenbaum and Katz (2008) led two independent panels to determine recommended cut scores on the Core and Advanced iSkills™ Assessments, which consists of 63 items within 15 performance-based tasks. In the end, the two independent panels converged on recommended scores (i.e., the mean of the cut scores of the two panels) corresponding to the core and intermediate foundational levels. To form smaller subpanels from a single panel is a popular and less costly variation on standard setting design; the legitimate uses of multivariate $(p^\bullet : s^\bullet) \times i^\circ$ and $(P^\bullet : S^\bullet) \times I^\circ$ designs² can certainly help to detect subpanel effect and to investigate generalizability over subpanels.

Under a multivariate G theory framework, the current study focuses on the modified Angoff method in standard setting. The Angoff family of procedures is one of the most popular and thoroughly researched of all currently used standard setting methods and will remain its widespread use in the future (Cizek & Bunch, 2007; Plake & Cizek, 2012). However, no method is free of criticisms and/or problems. Recently, the issues with the Angoff standard setting method with multiple-choice items have been well documented in Plake and Cizek (2012) to which readers of interest could refer. Major criticisms include the inability of the Angoff standard setters to form and maintain the integrated conceptualizations of proficiency required for implementing item-based procedures, and the incapability to making accurate item performance estimates (i.e., item ratings), especially for difficulty or easy items. Notwithstanding, the Angoff family is still attractive in practice due to its long history, the support from a massive body of literature, and some favoring features as the leading psychometricians with expertise in standard setting defended (see, for example, Hambleton et al., 2000). Not surprisingly, the Angoff family will continue its popularity in the future.

In conclusion, the utility of generalizability theory allows researchers to access the many sources of errors in standard setting. By the D study designs, the current study shows the impact of panel size and test length on measurement precision. For those who might replicate the modified Angoff procedures in a similar way, the study provides useful information for making decisions on desirable numbers of panelists and/or items such that the acceptable precision of cut scores is achieved. For those who might apply the other item- or test-centered standard setting methods, the study demonstrates a way of applying multivariate G theory designs when fixed effects exist. By inspecting the SE s of cut scores, the designers of standard setting agenda will be able to know whether the method delivered works properly. Finally, although the panelist

facet is random, a convenience sample has been used in collecting the standard setting data. Therefore, making clear arguments about generalization to a well-specified panelist population is relatively difficult, and one should be careful and cautious about that.

Appendix

Example Code of mGENOVA for Round 1 at Basic Level Using $p \times i^\circ$ Design

```
GSTUDY      Multivariate  $p \times i$  with Covariance Components Design =  $p$  Design: G4_R1_B
OPTIONS     NREC4 "*.out" TIME NOBANNER DEFAULT_DSTUDY
MULT        5 ALGEBRA STATISTICS GEOMETRY MEASUREMENT NUMBERSSEN
EFFECT      * $p$  6 6 6 6 6
EFFECT       $i$  11 10 10 30 40
FORMAT      0 4
PROCESS
*** 6 records with 101 observations per record ***
COMMENT End of G study
DSTUDY       $P \times I$  Design: D study sample sizes differ
DEFFECT     $  $P$  6 6 6 6 6
DEFFECT     1 5 5 15 15 20
ENDDSTUDY
FINISH
```

Authors' Note

An earlier version of this article was presented at the 2010 Annual Meeting of the American Educational Research Association.

Acknowledgments

The authors express their gratitude to the editor, Dr. Hua-Hua Chang, and three anonymous reviewers for their comments, which greatly improved the clarity and focus of the manuscript. They thank Dr. David Woodruff and Dr. Kate Hancock for editing the manuscript. They also thank Ms. Isabel Zheng and Mr. Edison Choe for their help in the process of publication. The first author is grateful to Dr. Stephen Dunbar, the Director of the Iowa Testing Programs, for his encouragement for this paper.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The conduction of the standard setting was part of the Taiwan Assessment of Student Achievement in Mathematics, funded by the National Academy for Educational Research, Taiwan.

Notes

1. The authors thanked one of the reviewers for correcting a misuse of the two-facet nested design in the former draft and for suggesting the use of two single-facet crossed designs in the current article due to the fact that the group effect in the study was fixed instead of random.
2. Readers interested in these designs are welcome to contact the authors.

References

- Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1997). *Technical report of the NAEP 1996 state assessment program in mathematics*. Washington, DC: National Center for Education Statistics.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- Arce-Ferrer, A., & Yin, P. (2007, April). *Standard errors of cut scores for vertically scaled assessments: A generalizability theory study of Angoff-based standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59-88.
- Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. In M. L. Bourque (Ed.) *Proceedings of the joint conference on standard setting for large-scale assessments* (Vol. II, pp. 269-287). Washington, DC: National Center for Education Statistics and National Assessment Governing Board.
- Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer-Verlag.
- Brennan, R. L. (2001b). *Manual for mGENOVA* (Iowa Testing Programs Occasional Paper No. 47). Iowa City: University of Iowa.
- 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.
- Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education, 12*, 151-165.
- Chang, L., & Hocevar, D. (2000). Models of generalizability theory in analyzing existing faculty evaluation data. *Applied Measurement in Education, 13*, 255-275.
- Cizek, G. J. (2012a). The forms and functions of evaluations of the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 165-178). New York, NY: Routledge.
- Cizek, G. J. (2012b). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 3-14). New York, NY: Routledge.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting—A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). 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.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process 1, 2. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., & . . . Zwick, R. (2000). A response to "setting reasonable and useful performance standards" in the National Academy of Sciences' grading the nations report card. *Educational Measurement: Issues and Practice, 19*(2), 5-14.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Washington, DC: American Council on Education.
- Kane, M., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8*, 107-115.
- Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., & van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education, 37*, 132-139.
- Lee, G., & Lewis, D. M. (2001, April). *A generalizability theory approach toward estimating standard errors of cutscores set using the bookmark standard setting procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

- Lee, G., & Lewis, D. M. (2008). A generalizability theory approach to standard error estimates for bookmark standard settings. *Educational and Psychological Measurement*, 68, 603-620.
- No Child Left Behind Act. (2002). Public Law 107-110 (20 U.S.C. 6311). Retrieved from <http://www2.ed.gov/policy/elsec/leg/esea02/index.html>
- Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 181-200). New York, NY: Routledge.
- Reckase, M. D. (1998). *Setting standards to be consistent with an IRT item calibration*. Iowa City, IA: ACT.
- Reckase, M. D. (2000, April). *The ACT/NAGB standard setting process: How "modified" does it have to be before it is no longer a modified-Angoff process?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Tannenbaum, R. J., & Katz, I. R. (2008). *Setting standards on the core and advanced iSkills™ assessments* (ETS RM-08-04). Princeton, NJ: Educational Testing Service.
- Tzou, H., Wu, Y.-F., & Lin, C. J. (2008, March). *Validating the performance standards in 2006 TASA-MAT standard setting*. Paper presented at the 2008 annual meeting of the National Council on Measurement in Education, New York City, NY.
- Verhoeven, B. H., van der Steeg, A. F. W., Scherpbier, A. J. J. A., Muijtjens, A. M. M., Verwijnen, G. M., & van der Vleuten, C. P. M. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education*, 33, 832-837.
- Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. *Medical Education*, 36, 860-867.
- Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 4, 663-704.
- Wu, Y.-F., & Tzou, H. (2010, April). *Evaluating the utility and validity of IRT-based approaches in the modified Angoff standard-setting method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.
- Yin, P., & Sconing, J. (2006, April). *Estimating standard errors of cut scores for Angoff-based and bookmark-based procedures: A generalizability theory approach*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- 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.