Journal of Early Adolescence
1–29
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0272431615588952
jea.sagepub.com
SAGE

# Assessing Mathematics Self-Efficacy: How Many Categories Do We Really Need?

Michael D. Toland[1] and Ellen L. Usher[1]

**Abstract**

The present study tested whether a reduced number of categories is optimal for assessing mathematics self-efficacy among middle school students using a 6-point Likert-type format or a 0- to 100-point format. Two independent samples of middle school adolescents ($N$ = 1,913) were administered a 24-item Middle School Mathematics Self-Efficacy Scale using either a 101-point or a 6-point response scale format. The findings suggest that the two different response scale formats were optimally represented by a 4-point scale and supported when samples were pooled. Results provide tentative evidence that middle school students make use of only 4 scale points and that the items on this scale are best matched with adolescents with average to below-average mathematics self-efficacy. Implications for the measurement of self-efficacy and related motivation constructs are discussed, and replications with a 4-point scale using category descriptors for each scale point are needed.

**Keywords**

mathematics self-efficacy, Rasch rating scale, response process evidence

[1] University of Kentucky, Lexington, USA

**Corresponding Author:**
Michael D. Toland, Department of Educational, School, and Counseling Psychology, University of Kentucky, Lexington, KY 40506, USA.
Email: toland.md@uky.edu

Over the past three decades, researchers have consistently shown that the belief students hold in their mathematics capabilities, or *mathematics self-efficacy*, is predictive of students' achievement and motivation in mathematics, such as their goal orientation, value, and self-concept (Brown & Lent, 2006; Cheema & Kitsantas, 2013; Hackett & Betz, 1989; Schunk & Pajares, 2005). Measures used to assess mathematics self-efficacy have varied in content, scope, and structure, but almost all have used a Likert-type response scale on which students are asked to indicate their level of certainty that they can accomplish a given task. Self-efficacy measures that follow guidelines put forth by Bandura (2006) ask respondents to use a 0 to 100 response scale to provide judgments of their capability to successfully perform various tasks, although many researchers have opted for response scales with fewer categories. Some researchers have questioned whether a 0 to 100 response scale format is ideal when measuring self-efficacy. Moreover, both developmental psychologists and psychometricians have noted that younger respondents in particular might find too many response categories cognitively overwhelming (e.g., Cowan, 2010) or beyond their grasp to differentiate metacognitively (e.g., Weil et al., 2013). We address this question empirically and offer recommendations for the measurement of self-efficacy and related motivation constructs.

First, we provide a brief synopsis of how students' self-efficacy is assessed in the domain of mathematics. We then review research aimed at investigating the appropriate response format for self-efficacy measures. Next, we offer an empirical analysis of the utility of a 6-point versus a 0- to 100-point response scale designed to assess middle school students' self-efficacy for performing mathematics skills. We discuss findings as they are situated in the field of educational psychology, in which researchers often rely on similar self-report response scales to explain psychological phenomena. Finally, implications for educational and social science researchers and practitioners are offered.

## Assessment of Mathematics Self-Efficacy

Researchers interested in self-efficacy as an explanatory construct must first identify and define what it means to be efficacious, or competent, in a given domain (e.g., to be able multiply fractions). Items to assess beliefs in one's efficacy can then be designed with this criterion in mind. Statements or questions are then given to respondents who in turn rate their sense of efficacy for achieving the stated benchmark (e.g., "How confident are you that you can *multiply fractions*?"). When self-efficacy and achievement measures closely correspond, the predictive power of self-efficacy is enhanced (Bong, 2001; Pajares & Miller, 1995). For example, Pajares and Barich (2005) found that a measure of students' self-efficacy for earning a particular grade in high school mathematics was a better predictor of actual grades earned than was a measure of students' mathematics skills self-efficacy. A host of research has shown that domain-specific self-efficacy measures (e.g., mathematics self-efficacy) are strong positive predictors of students' academic behaviors and subsequent motivation in the same domain (see Klassen & Usher, 2010). Generalized self-efficacy measures that are not specific to a domain are typically unrelated to these same outcomes.

Researchers have not reached consensus on how many response categories to use when assessing students' self-efficacy. Just as the level of specificity in self-efficacy assessment has varied, so has the response format used across studies. Researchers have largely relied on Likert-type response formats that vary in range from 4 to 10 category systems (e.g., Morony, Kleitman, Lee, & Stankov, 2013; Pajares & Graham, 1999; Pintrich & De Groot, 1990; Usher & Pajares, 2009). Others have followed Bandura's (2006) guidelines by utilizing a 101-point response format in which students are asked to write in a number from 0 to 100 to indicate their level of confidence. Some have opted for a variation of these two approaches by asking students to rate their confidence on a scale ranging from 0 to 100 in 10-unit intervals (e.g., Bong & Hocevar, 2002). Labels on self-efficacy response scales have also varied. On most self-efficacy scales, anchors are provided at scale endpoints (e.g., *not confident at all* to *completely confident*) or at key interval points (e.g., 1 = *not at all true*, 3 = *somewhat true*, and 5 = *very true*). Rarely do researchers provide anchors or word labels at every possible response location, even though a word label for each response category would ensure that everyone uses the same response category labels as a means of reducing measurement error (DeVellis, 2012).

## Research on Response Format in Self-Efficacy Measures

Some attempts have been made by researchers to determine the optimal response format for self-efficacy measures in terms of its influence on reliability and validity. In one such study, Maurer and Pierce (1998) compared a 5-point, Likert-type format with a two-pronged assessment in which college students were asked first to indicate whether they could perform academic self-regulation tasks at a given level (i.e., yes or no) and then to rate their confidence in that assessment (i.e., 0-100). They found the Likert-type format to be a better alternative to the latter method with respect to classical reliability estimates (α), levels of prediction, factor structure, and discriminability.

In a similar study, Pajares, Hartley, and Valiante (2001) compared a 6-point, Likert-type self-efficacy scale with a 0 to 100 self-efficacy scale in the context of writing in middle school. They found that the 0 to 100 scale provided similar results to the Likert-type format with respect to classical reliability estimates and factor structure, but the 0 to 100 format provided psychometrically stronger evidence with respect to levels of prediction/correlation and discriminability. Kan (2009) showed that a teaching self-efficacy scale with a 0 to 100 format had relatively better psychometric properties than did either a 0 to 100 visual analog scale or a 6-point Likert-type confidence scale with respect to classical reliability, explained percentage of total variance, size of pattern loadings, levels of prediction, and generalizability coefficient.

Notable limitations exist in the research aimed at comparing the psychometric properties of these response formats, however. Most researchers have used classical test theory or factor analytic approaches to examine response formats and patterns. These approaches are typically used to evaluate self-efficacy measures in other domains (see Klassen & Usher, 2010). For example, a review of self-efficacy instrumentation used in the domain of writing revealed that classical test theory or factor analysis were used in 96% of studies (Butz, Toland, Zumbrunn, Danner, & Usher, 2014). In only 2 of 50 studies did researchers report using Rasch or item response theory (IRT) approaches.

Why might the measurement approach matter when it comes to investigating learners' self-efficacy? First, a classical test theory approach assumes that the interval between the numeric points on each self-efficacy scale are linear or equidistant, and that scale scores therefore represent an individual's relative position in a score distribution. Such approaches involve the creation of a sum score or mean score for each student on a given scale and are frequently used in psychological and educational research. However, such mathematical operations can lead to spurious conclusions. For instance, Embretson (1996) showed that 2 × 2 interactions from ANOVA designs had inflated Type I error (false positive) rates when the outcome variable was derived from the summation of dichotomously scored items relative to items scaled with the Rasch (one-parameter logistic response) model. Results from a separate investigation indicated that interaction effects among continuous predictors were inflated when the outcome variable was derived from the summation of dichotomously scored items scaled with the two-parameter logistic response model (Kang & Waller, 2005). In both studies, the spurious interaction effect was most evident when test difficulty was poorly matched with the sample characteristics. In other words, a spurious effect could be found when self-efficacy items represent lower skill levels (i.e., items are easy for highly skilled respondents to endorse). The spurious effect was not present when analyses were conducted with Rasch or IRT estimated trait scores (Kang & Waller, 2005). Researchers have also shown that scales that use beyond 4 to 7 response points offer few gains in terms of classical reliability and validity estimates (Lozano, Garciá-Cueto, & Muñiz, 2008).

A second consideration for researchers wishing to determine optimal response formats is the cognitive complexity of the judgments individuals are asked to make. Respondents' ability to make fine-tuned judgments of their own efficacy might be directly related to their expertise in a given domain. For example, a student taking advanced algebra who understands the complexity of algebraic equations may not hesitate to provide a 60 out of 100 rating of her efficacy to solve an equation with two variables; a mathematical neophyte unfamiliar with the steps involved in solving for two variables may select a 50 out of 100 rating as a middle-of-the-road estimate of what he could learn to do. In this case, providing too many response categories could lead to variance due to method rather than due to underlying differences in content-related self-efficacy. Respondents, particularly novice learners, may have trouble differentiating between the many responses possible on a 0 to 100 scale, which may invoke an undue cognitive burden and lead to limited use of response categories. Indeed, some cognitive psychologists have argued that people in general, and young learners in particular, are not capable of holding more than several categories in working memory at a time (Cowan, Morey, Chen, Gilchrist, & Saults, 2008). During late childhood and adolescence, students' metacognitive skills, which include judgments about one's own capabilities, are in development and may not be fine-tuned enough to discriminate between tens (or hundreds) of information categories (Schneider, 2008).

Researchers naturally decide on the number of response categories to use on a given multi-item measure prior to administering it to participants. After data have been collected, however, a simple inspection of the response frequencies does not provide evidence that the ordering of the rating scale categories has been used by participants in the intended way. Rasch or IRT measurement techniques can provide researchers with an empirical means for evaluating how participants use the rating scale categories. This type of validity evidence is known as response process evidence and addresses the fit of the construct being studied with the actual responses observed by respondents (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Readers are referred to de Ayala (2009) for more details on Rasch and IRT techniques.

Several researchers have used Rasch approaches in the measurement of self-efficacy. For example, E. V. Smith, Wakely, De Kruif, and Swartz (2003) used a Rasch rating scale model to investigate a writing self-efficacy scale for upper elementary students. This modeling technique revealed that respondents who were presented with a 10-point scale ranging from 10 to 100 (in increments of 10) with four anchors actually made use of only four primary categories. This finding was replicated by the researchers in a second independent sample. A Rasch rating scale model was also used to examine postsecondary students' mathematics self-efficacy, which was assessed on a 4-point scale, but the researchers did not use the rating scale model to examine response category use.

Rasch models have also been applied in other domains and with other populations. For example, a Rasch rating scale model has been used to examine the optimal number of rating scale categories for measuring caregiver self-efficacy (Cipriani, Hensen, McPeck, Kubec, & Thomas, 2012), and nursing self-efficacy (Hagquist, Bruce, & Gustavsson, 2009). Findings from these studies suggest that a reduced-category system worked better than an 11-point scale ranging from 0 to 100 (in increments of 10). It bears restating, however, that these measurement studies have been based on data gathered in different (largely nonacademic) contexts and with older participants, and may not generalize to other age groups and populations (e.g., students in middle school mathematics). These findings may also not be similar to those obtained from different instruments, such as response scales that start with fewer response points (e.g., 6 points) or those that allow respondents to provide responses to the nearest integer from 0 to 100.

The purpose of the present study was to use the Rasch rating scale model to test whether a reduced number of categories is optimal for assessing mathematics self-efficacy among middle school students using a 6-point Likert-type format or a 0- to 100-point format. Based on previous self-efficacy research and other domains and samples, we hypothesized that a reduced-category system would be more appropriate for middle school students. The Rasch method was elected to address limitations inherent in classical approaches (i.e., assumed response order) that have been most often used in self-efficacy research. As Andrich (2013) indicated, a Rasch model can be used to assess a hypothesis about item category order.

## Method

### Participants

Data were collected in November 2006 as part of a larger study involving two independent samples of students attending two middle schools from a school district in the Southeastern United States. The samples were not drawn randomly from a population but were convenience samples from two schools. Each school was considered as a separate sample in our study; thus, potential implications may be sample specific until further replication. Sample 1 consisted of 1,110 students (372 sixth, 375 seventh, and 363 eighth graders) with an average age of 12.27 years (*SD* = 0.93), who were 50.4% female and self-identified as White (61.0%), Asian (16.9%), African American (12.3%), Hispanic (5.2%), and Other ethnicity (2.4%). Sample 2 consisted of 803 students (282 sixth, 255 seventh, and 266 eighth graders) with an average age of 12.22 years *SD* = 0.94), who were 50.8% female and self-identified as White (67.4%), African American (18.7%), Hispanic (6.4%), Asian (3.5%), and Other ethnicity (4.1%).

## Instrument and Procedure

The second author administered a 24-item Mathematics Skills Self-Efficacy Scale to students in intact mathematics classrooms at a time convenient to teachers. This scale was part of a larger survey on mathematics motivation. Items were crafted to reflect the middle school mathematics learning standards (e.g., use of ratios and proportions) of the National Council of Teachers of Mathematics (NCTM; 2000) and in accordance with guidelines for constructing self-efficacy items (Bandura, 2006; Bong, 2006). Students were asked to rate how confident they are at successfully completing exercises related to 24 mathematics topics without using a calculator on either a scale with descriptive anchors at 1 (*not at all confident*) and 6 (*completely confident*) or a 0 to 100 confidence scale with descriptive anchors at 0 (*not at all confident*), 50 (*somewhat confident*), and 100 (*completely confident*). Items on both surveys were placed on one page with one sentence as the stem (i.e., "How confident are you that you can successfully solve mathematics exercises involving . . . ") and 24 statements completing the stem (e.g., order of operations?; copies of the instruments along with each item are provided in Appendices A and B). We opted to use general mathematics topics rather than providing specific mathematics problems as this corresponded to the outcome of interest in the study (i.e., course grade in mathematics).

Students in Sample 1 received the 1 to 6 rating scale and were asked to circle the number that corresponded to their confidence level. Students in Sample 2 received the 0 to 100 confidence scale and were allowed to write in any number between 0 and 100. Students completed the form independently and could ask the researcher questions at any time. Regardless of format, each student took approximately 5 min to complete the scale.

## Data Analysis

The data were analyzed using the Rasch rating scale model (Andrich, 1978), which is appropriate for polytomous data, via the Winsteps 3.72 program (Linacre, 2011). Although a continuous rating scale model (Müller, 1987) could be applied to the 0 to 100 scale instead of the rating scale model, we did not view the continuous model appropriate for adding value to the current study. Moreover, studies have found close agreement between the two models' results (see Eckes, 2011). We also wanted to be consistent in the model applied to both rating scales.

An underlying assumption of the Rasch rating scale measurement model (Andrich, 1978) is unidimensionality; therefore, a principal components analysis of the standardized residuals (PCAR) on the final scale in each sample and pooled sample was performed via Winsteps 3.72 (Linacre, 2011). In a Rasch analysis, the first dimension is the Rasch measurement model imposed on the data. So, the first component of the PCAR is the largest secondary dimension (first contrast). Essential unidimensionality was considered by examining the variance explained by the Rasch dimension, unexplained variance by the secondary dimension (i.e., first contrast or size of eigenvalue), inspection of items at top and bottom of standardized residual contrast 1 plot, and correlating adolescent measures (i.e., two subsets of scores were based on positive and negative item loadings on the first residual dimension).

Once the data were fit to the Rasch model, infit and outfit item statistics were evaluated and items removed if outside of the 0.5 to 1.5 range, with statistics centered around 1 (Linacre, 2009; R. M. Smith, 1996, 2000; Wright & Stone, 1999). The item fit statistics indicate whether the items are performing in a manner consistent with the rating scale model. Person reliability (i.e., "a ratio of variance adjusted for measurement error to observed variance, conceptually equivalent to Cronbach's alpha" E. V. Smith et al., 2003, p. 378), item reliability (i.e., a measure that indicates how spread out items are along the mathematics self-efficacy continuum, similar to Cronbach's coefficient α), person separation (i.e., the degree to which an instrument can separate apart persons with varying levels of latent variable; de Ayala, 2009, p. 54), and item separation (i.e., the degree to which an instrument can separate apart items along the latent variable continuum; de Ayala, 2009, p. 55) were also reported along with an item-person map to summarize the discrepancy between item and person estimates. A category probability curve was reported for the final rating scale structure to demonstrate the probability of selecting a particular response category given one's level of mathematics self-efficacy. This probability is expressed in logits.

To determine the final rating scale structure, we used guidelines set forth in E. V. Smith et al. (2003) and Linacre (2002): Each category has at least 10 observations, regular observation frequency are observed across categories, average measures are ordered or advance monotonically with category, outfit mean-square statistic is less than 1.4 (i.e., "a high mean-square associated with a particular category indicates that the category has been used in unexpected contexts"; Linacre, 2002, p. 93), and thresholds or step calibrations are ordered. As recommended by Andrich (2013), both conceptual and empirical evidence were used to guide decision making. We collapsed categories based on substantive reasoning or on responses which signified the same level of mathematics self-efficacy. We used the threshold order derived from the Rasch rating scale model to flag any anomalies (disordering), and we collapsed categories that were substantively similar. Once the resulting optimal response format for both samples was found, further analysis was performed with the two samples pooled together by fitting the rating scale model to the data and inspecting fit, conducting differential item functioning analyses within Winsteps, and reporting item-level descriptive statistics for the final rating scale.

## Results

### Dimensionality Assessment

Initial dimensionality results for the 24-item, 6-point scale in Sample 1 and the 0- to 100-point scale in Sample 2, indicated that the primary dimension explained 45.7% (eigenvalue = 20.2) and 45.1% (eigenvalue = 21.1) of the raw variance, and the first residual dimension accounted for 5.2% (eigenvalue = 2.3) and 5.7% (eigenvalue = 2.6) of the unexplained variance in the sample, respectively. An inspection of the items at the top and bottom of a standardized residual contrast 1 plot (not reported) did not show substantive differences in either sample. Also, two subsets of items were created by splitting the items based on positive and negative loadings on the first residual dimension and examining correlations among the adolescent measures (scores). The scores were correlated at .86 in each sample. Overall, these results suggest that the scale in each sample was substantively unidimensional (Linacre, 2003). As expected, after optimizing the number of response categories in each sample and using the pooled sample (i.e., Samples 1 and 2 combined), the conclusion was the same. In the following sections, we present results that led to the final optimized rating scale.

### Rating Scale Results for Sample 1

Initial results from the Rasch rating scale analysis for the 24-item, 6-point scale indicated person and item reliabilities of .85 and .99, with person and

**Table 1.** Sample 1 ($n$ = 1,110) Initial Category Counts, Average Measures, Threshold (Structure, Step) Measures, and Outfit Mean-Square Statistics for a 6-Point Scale.

| Category label | Observed count | Average measure | Outfit $\chi^2$ | Threshold |
|---|---|---|---|---|
| 1 | 638 | −0.28 | 1.37 | |
| 2 | 740 | 0.02 | 1.19 | −0.44 |
| 3 | 1,810 | 0.32 | 0.97 | −0.76 |
| 4 | 3,445 | 0.71 | 0.88 | −0.12 |
| 5 | 6,514 | 1.24 | 0.81 | 0.35 |
| 6 | 13,483 | 2.05 | 1.04 | 0.96 |

item separations of 2.36 and 13.59. Higher separation values are better, but there is no clear cutoff value. A summary of the observed category counts, average measures, thresholds, and outfit mean-square statistics are provided in Table 1. For the 6-point scale, our criteria for observed count were met, average measures demonstrated monotonicity (i.e., higher categories manifest higher self-efficacy levels than lower categories), and all outfit mean-square statistics were less than 1.4. However, thresholds were disordered. This disordering or reversal means that "a given category is not as likely to be chosen as the other categories" (de Ayala, 2009, p. 193). In terms of our data, as early adolescents with higher levels of mathematics self-efficacy were observed, each new self-efficacy rating scale category was not more likely to be chosen than previous categories, as would be expected. This same information was also found in a category probability curve or option response function (not reported). Thus, the rating scale structure did not operate optimally or as expected, which jeopardizes response process validity.

Given the reversals in the thresholds occurring between Categories 2 and 3 (see Table 1), categories were combined that were not performing as expected (i.e., disordered) with adjacent categories below them. Therefore, the original 6-point category codes of 123456 were collapsed into a 5-point category coding system of 122345. This latter expression meant the original category code of 1 was retained, original category codes of 2 and 3 were changed to 2, original code 4 was changed to 3, original code 5 was changed to 4, and original code 6 was changed to 5. The collapsing of original category codes 2 and 3 into a single category was deemed substantively sensible given that both codes occur immediately after the first category (*not at all confident*), but below the top three codes that suggest more or above average confidence.

Analysis of this 24-item, 5-point scale showed a slight increase in person reliability (.86), item reliability stayed the same (.99), and person and item

**Table 2.** Sample 1 (*n* = 1,110) Category Counts, Average Measures, Threshold (Structure, Step) Measures, and Outfit Mean-Square Statistics for a 5-Point Scale.

| Category label | Observed count | Average measure | Outfit $\chi^2$ | Threshold |
|---|---|---|---|---|
| 1 | 638 | −0.35 | 1.17 | |
| 2 | 2,550 | 0.20 | 1.04 | −1.56 |
| 3 | 3,445 | 0.72 | 0.93 | 0.14 |
| 4 | 6,514 | 1.30 | 0.82 | 0.37 |
| 5 | 13,483 | 2.16 | 1.06 | 1.05 |

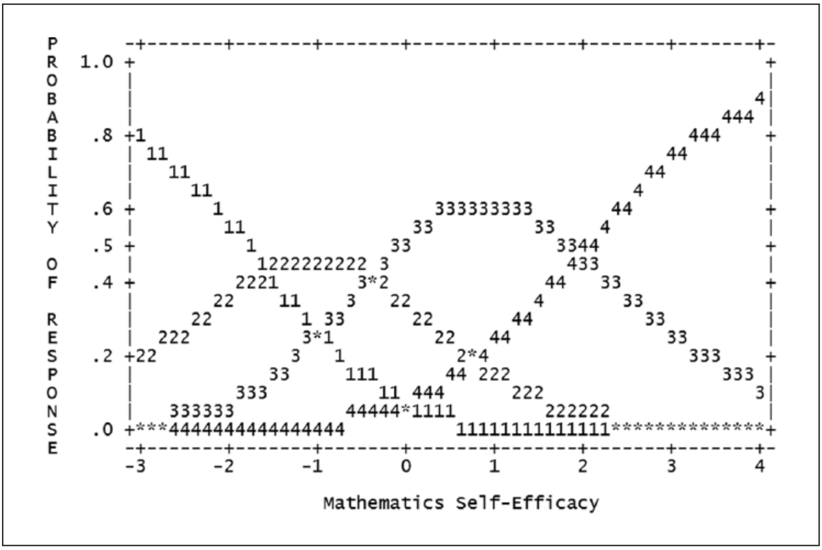
separations increased slightly to 2.45 and 13.78. Evaluation of the 5-point scale results showed observed category counts were met, average measures were ordered, thresholds were ordered, and outfit mean-square statistics were all less than 1.4 (see Table 2). However, inspection of the change in thresholds for category labels 3 and 4 showed minimal changes (0.37-0.14 = 0.23 logits), which is below Linacre's (2002) suggestion that thresholds change by at least 1.4 logits. Also, each category in a probability curve should have a discernable peak or hill and smooth and a discernable step between categories. The step between Categories 3 and 4 was barely discernable in a category probability curve (not reported) or barely showed a distinct hill, suggesting that Categories 3 and 4 should be combined. According to the recommendations by Linacre (2002), this rating scale structure was not optimal.

Based on these results, we deemed it substantively reasonable to collapse the two middle-high category codes of 3 and 4 into a single code. The 4-point category coding system was 122334. Substantively, the four-category coding system can be deemed meaningful because early adolescents could view the bottom two middle categories similarly, and the two middle-high categories similarly, while viewing the extreme category codes as distinct categories.

Analysis of the 24-item, 4-point scale indicated a slight increase in person reliability (.88), item reliability stayed the same (.99), person separation increased slightly to 2.65, but item separation had a slight decrease (13.30). Evaluation of the 4-point scale structure indicated that observed category counts were met, average measures were ordered, thresholds were ordered, and outfit mean-square statistics were all less than 1.4 (see Table 3). The 4-point rating scale structure was also demonstrated in the category probability curves in Figure 1. Specifically, adolescents less than −1.56 logits on the mathematics self-efficacy continuum had higher probabilities of selecting Category 1, adolescents between −1.56 and −0.4 logits were more likely to select Category 2, adolescents between −0.4 and 1.96 logits were more likely to choose Category

**Table 3.** Sample 1 (*n* = 1,110) Category Counts, Average Measures, Threshold (Structure, Step) Measures, and Outfit Mean-Square Statistics for a 4-Point Scale.

| Category label | Observed count | Average measure | Outfit $\chi^2$ | Threshold |
|---|---|---|---|---|
| 1 | 638 | −0.44 | 1.23 | |
| 2 | 2,550 | 0.36 | 0.95 | −1.56 |
| 3 | 9,959 | 1.52 | 0.87 | −0.40 |
| 4 | 13,483 | 2.94 | 1.03 | 1.96 |



```
P      -+-------+-------+-------+-------+-------+-------+-------+-
R  1.0 +                                                         +
O      |                                                         |
B      |                                                        4|
A      |                                                     444 |
B   .8 +1                                                 444    +
I      | 11                                             44       |
L      |   11                                         44         |
I      |     11                                      4           |
T   .6 +       1                   333333333       44            +
Y      |        11               33         33    4              |
    .5 +          1            33             3344               +
O      |           1222222222 3                433               |
F   .4 +         2221       3*2              44   33             +
       |       22    11    3   22           4       33           |
R      |     22        1 33      22       44          33         |
E      |  222          3*1         22   44              33       |
S   .2 +22            3   1          2*4                  333    +
P      |            33     111      44 222                   333 |
O      |         333          11 444      222                   3|
N      |   333333          44444*1111        222222              |
S   .0 +***4444444444444444          11111111111111**************+
E      -+-------+-------+-------+-------+-------+-------+-------+-
       -3      -2      -1       0       1       2       3       4

                         Mathematics Self-Efficacy
```

**Figure 1.** Final category probability curves for a 4-point scale that was initially a 6-point scale.

3, and adolescents above 1.96 logits on the self-efficacy continuum were more likely to select Category 4. Note that a 3-point category scale was considered (i.e., whereby the category coding system was 112233), but this resulted in a loss of both person and item separability and relatedly reliability.

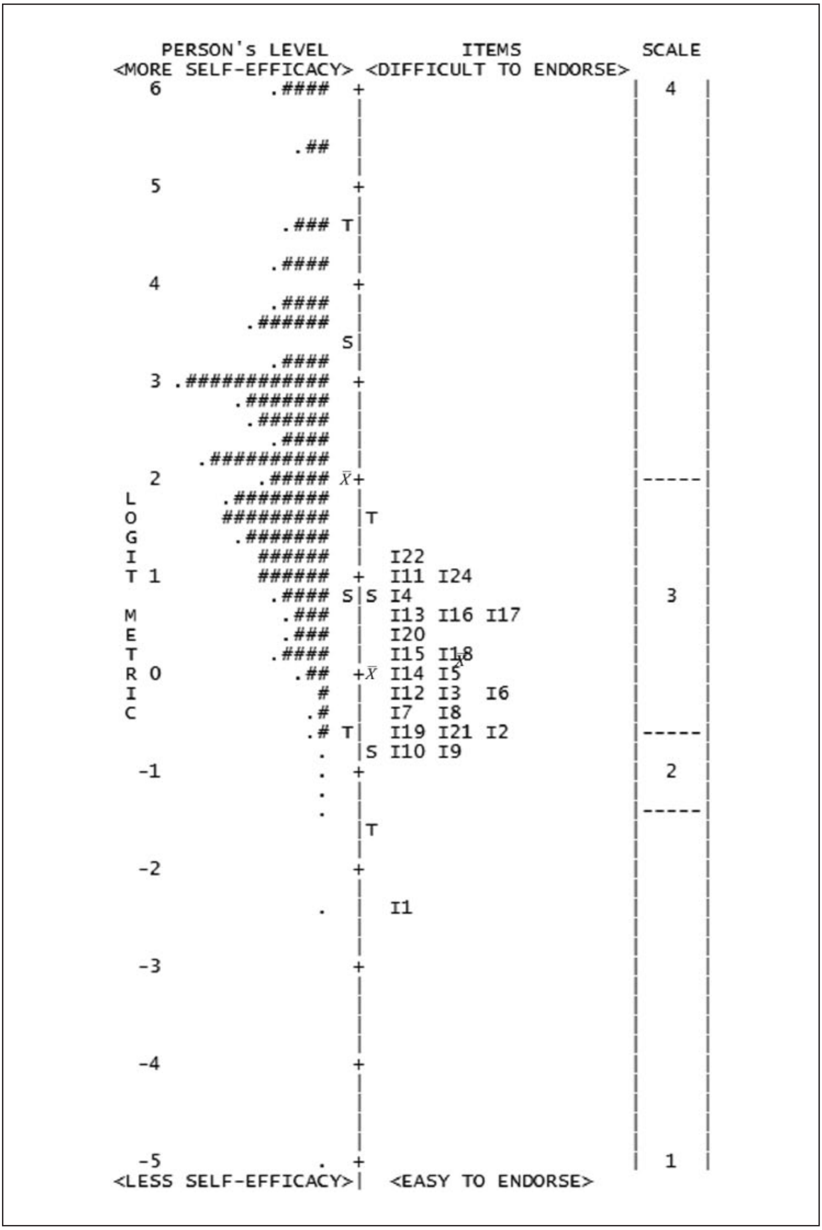
Using the 4-point scale, infit and outfit item statistics for all items were found to fall within the recommended range of 0.5 to 1.5 ( $\bar{X}$ ± *SD* of Infit statistics = 1.02 ± 0.15, Minimum = 0.71, Maximum = 1.47; $\bar{X}$ ± *SD* of Outfit statistics = 0.97 ± 0.16, Minimum = 0.62, Maximum = 1.41). This finding also gives evidence for the embedded assumption of a unidimensional structure in the Rasch rating scale model.

A variable map of the distribution of mathematics self-efficacy items and adolescents' mathematics self-efficacy level for a 4-point scale (initially a 6-point scale) is illustrated in Figure 2. The logit scale is shown on the far left of the map. The histogram in the middle-left of the map shows the distribution of students' mathematics self-efficacy level. The middle-right of the map shows each item at its respective difficulty-of-endorsement level. The map shows that the distribution of items ($\bar{X}$ ± *SD* of logit: 0 ± 0.7) is not well-targeted (statistically inappropriate) to the adolescents' mathematics self-efficacy level ($\bar{X}$ ± *SD* of logit: 2.19 ± 1.55), as demonstrated by the mismatch between person and item distributions.

The item hierarchies demonstrate how the adolescents perceived their efficacy in mathematics. The higher the logit score for a particular item, the more difficult it was for adolescents to endorse the item (i.e., to indicate that they were confident about their skills related to a given math topic), whereas the lower the logit score for an item, the easier it was for adolescents to endorse an item. For example, adolescents found it most difficult to endorse the following items: I4 (ratios and proportions), I11 (rounding and estimating), I22 (explaining in words how you solved a math problem), and I24 (doing quick calculations in your head). By contrast, they found it easier to endorse these items: I1 (multiplication and division), I9 (order of operations), and I10 (rounding and estimating). In general, items tended to target adolescents with average to below-average self-efficacy based on items falling at or below the mean for adolescents. In other words, because the mean self-efficacy score for the adolescents is greater than the mean of the items, it can be concluded that adolescents had an easy time stating that they are confident with the math topics. The category thresholds are depicted on the far right side of the map and indicate that the category thresholds increased across the rating scale.

## Rating Scale Results for Sample 2

Initial results for the 24-item, 0- to 100-point scale indicated person and item reliabilities of .24 and .99, with person and item separations of 0.56 and 8.48, respectively. Low item reliability and separation statistics indicate that the items were not spread out along the mathematics self-efficacy scale. Even though the metric affords a possible 101 discrete categories and the 24-item scale had an observed range from 0 to 99, initial counts were well below 10 for more than half the categories and modal frequencies tended to occur at roughly every 5th or 10th position. Therefore, rarely observed categories were combined with adjacent and substantively similarly meaningful categories to produce more stable thresholds.



```
     PERSON's LEVEL            ITEMS               SCALE
<MORE SELF-EFFICACY> <DIFFICULT TO ENDORSE>
    6          .####  +                          |  4   |
                      |                          |      |
                      |                          |      |
                 .##  |                          |      |
                      |                          |      |
    5                 +                          |      |
                      |                          |      |
                .### T|                          |      |
                      |                          |      |
               .####  |                          |      |
    4                 +                          |      |
               .####  |                          |      |
            .######   |                          |      |
                     S|                          |      |
               .####  |                          |      |
    3 .############   +                          |      |
           .#######   |                          |      |
            .######   |                          |      |
               .####  |                          |      |
        .##########   |                          |      |
    2       .##### X̄+                            |----- |
 L        .########   |                          |      |
 O        #########   |T                         |      |
 G         .#######   |                          |      |
 I           ######   |   I22                    |      |
 T  1        ######   +   I11 I24                |      |
              .#### S|S   I4                     |  3   |
 M             .###   |   I13 I16 I17            |      |
 E             .###   |   I20                    |      |
 T            .####   |   I15 I18                |      |
 R  0           .##  +X̄   I14 I5                 |      |
 I                #   |   I12 I3  I6             |      |
 C               .#   |   I7  I8                 |      |
                 .# T|    I19 I21 I2             |----- |
                  .   |S  I10 I9                 |      |
   -1             .   +                          |  2   |
                  .   |                          |      |
                  .   |                          |----- |
                      |T                         |      |
                      |                          |      |
   -2                 +                          |      |
                  .   |   I1                     |      |
                      |                          |      |
                      |                          |      |
   -3                 +                          |      |
                      |                          |      |
                      |                          |      |
                      |                          |      |
                      |                          |      |
   -4                 +                          |      |
                      |                          |      |
                      |                          |      |
                      |                          |      |
                      |                          |      |
   -5             .   +                          |  1   |
<LESS SELF-EFFICACY>|    <EASY TO ENDORSE>
```

**Figure 2.** Variable map for a 4-point scale that was initially a 6-point scale.
*Note.* # = 8 adolescents; . = 1 to 7 adolescents; $\overline{X}$ = mean; S = one standard deviation; T = two standard deviations; I = item; —— = threshold.

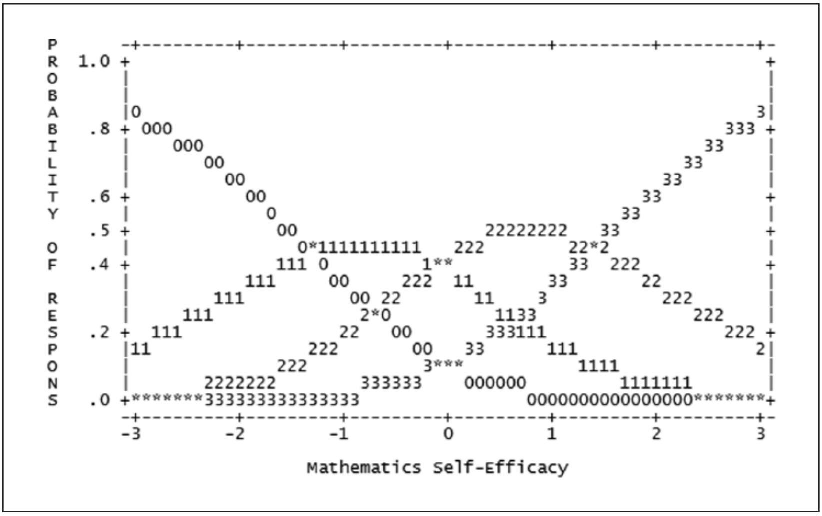
We used several criteria based on Wright's (1987) empirical rules (as cited in Wright & Linacre, 1992) for combining adjacent categories for a unidimensional, polytomous Rasch model. First, categories were combined that made substantive sense or signified the same level of mathematics-self efficacy. Second, only low frequency categories were combined with modal categories, as combining low frequency categories with other low frequency categories would have artificially created more modal categories and distorted results. Third, categories were combined upwards toward the highest category (i.e., any category with frequency less than 10 was combined with the next highest modal category) given that the items were seen as easy for the sample. In this way, the item category frequency profile matched the relevant segment of the sample of adolescents. Based on the low category counts in the 0 to 100 scale, responses were combined into a 0 to 42 scale (not reported).

Several intermediate analyses were conducted after this first combination of categories to arrive at the final rating scale structure. The number of scale points was reduced in each analysis in the following manner: the initial 0 to 100 scale was reduced to a 0 to 42 scale, then to a 0 to 15 scale, next a 0 to 13 scale, a 0 to 12 scale, a 0 to 11 scale, a 0 to 4 scale, and lastly to a 0 to 3 scale. The first revision occurred because of low category counts, the next four revisions occurred because of disordered average measures, and the final two revisions occurred because of disordered thresholds. However, combining of categories was ultimately done when collapsing could be substantiated. Person reliability increased after each analysis, item reliability stayed the same (.99), person separation slowly increased after each analysis from 0.56 (0-100 scale) to 2.78 (0-3 scale), and item separation fluctuated between 8.48 (0-100 scale) and 9.65 (0-15 scale). Note, a 0- to 2- or 3-category scale was considered, but this resulted in a loss of both person and item separability and reliability. The final 4-point category coding system was 0 to 35 = 0, 36 to 70 = 1, 71 to 90 = 2, and 91 to 100 = 3. Readers interested in how the results changed from 0 to 42 categories and to some of the smaller ones can email the first author.

Analysis of the 24-item, 4-point scale (i.e., the 0-3 scale) showed a person reliability of .89, item reliability of .99, item separation of 8.90, and person separation of 2.78. Evaluation of the 4-point scale results showed that observed category counts were met, average measures were ordered, thresholds were ordered, and outfit mean-square statistics were less than 1.4 (see Table 4). The 4-point rating scale structure is illustrated in the category probability curves in Figure 3. Specifically, adolescents who were lower than −1.30 logits on the mathematics self-efficacy continuum had higher probabilities of selecting Category 0, adolescents between −1.30 and −0.10 logits were more likely to select Category 1, adolescents between −0.10 and 1.40 were more likely to choose Category 2, and adolescents above 1.40 logits on the self-efficacy continuum were more likely to select Category 3.

**Table 4.** Sample 2 (n = 803) Category Counts, Average Measures, Threshold (Structure, Step) Measures, and Outfit Mean-Square Statistics for a 4-Point Scale.

| Category label | Observed count | Average measure | Outfit $\chi^2$ | Threshold |
|---|---|---|---|---|
| 1 | 1,315 | −0.67 | 1.07 | |
| 2 | 3,575 | 0.09 | 0.91 | −1.30 |
| 3 | 6,741 | 0.97 | 0.89 | −0.10 |
| 4 | 7,641 | 1.93 | 1.09 | 1.40 |



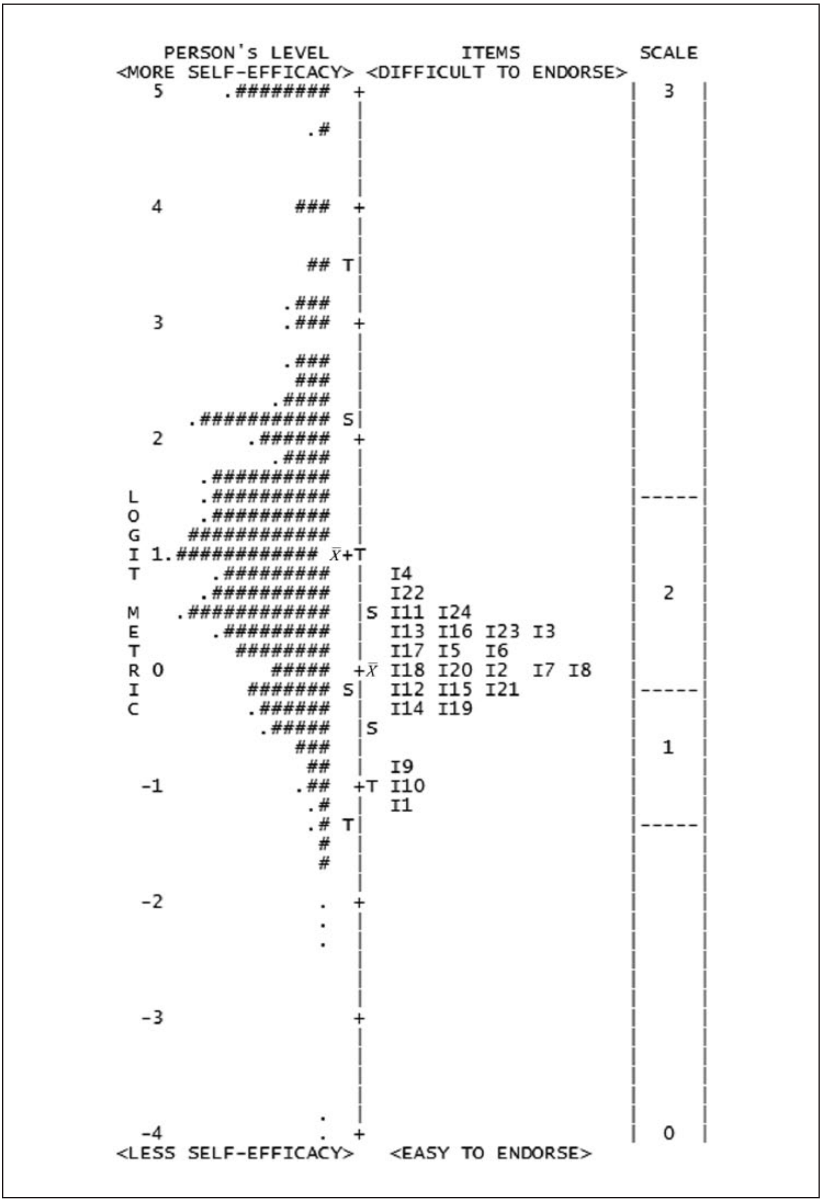
```
P      -+---------+---------+---------+---------+---------+---------+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |0                                                           3|
B   .8 + 000                                                     333 +
I      |    000                                                33    |
L      |       00                                            33      |
I      |         00                                        33        |
T   .6 +           00                                    33          +
Y      |             0                                 33            |
    .5 +              00                  22222222   33              +
O      |                0*1111111111   222        22*2               |
F   .4 +              111 0         1**           33  222            +
       |           111     00     222  11       33       22          |
R      |        111          00 22       11    3           222       |
E      |     111              2*0          1133               222    |
S   .2 +  111               22   00       333111                 222 +
P      |11               222       00   33      111                 2|
O      |              222           3***           1111              |
N      |       2222222        333333    000000         1111111       |
S   .0 +*******333333333333333                0000000000000000*******+
       -+---------+---------+---------+---------+---------+---------+-
       -3        -2        -1         0         1         2         3

                         Mathematics Self-Efficacy
```

**Figure 3.** Final category probability curves for a 4-point scale that was initially a 0 to 100 scale.

Examination of the 4-point scale revealed that infit and outfit item statistics for all items fell within the recommended range of 0.5 to 1.5 ( $\bar{X}$ ± *SD* of Infit statistics = 1.01 ± 0.16, Minimum = 0.68, Maximum = 1.49; $\bar{X}$ ± *SD* of Outfit statistics = 0.99 ± 0.15, Minimum = 0.69, Maximum = 1.45). This finding also provides evidence for the embedded assumption of a unidimensional structure in the Rasch rating scale model.

A variable map of the distribution of mathematics self-efficacy items and adolescents' mathematics self-efficacy level for a 4-point scale (initially a 0- to 100-point scale) is illustrated in Figure 4. The map shows that the distribution of items ( $\bar{X}$ ± *SD* of logit: 0 ± 0.49) is not well-targeted (statistically inappropriate) to the adolescents' mathematics self-efficacy level ( $\bar{X}$ ± *SD*



```
    PERSON's LEVEL           ITEMS          SCALE
<MORE SELF-EFFICACY> <DIFFICULT TO ENDORSE>
   5     .########  +                      |  3  |
                    |                      |     |
                .#  |                      |     |
                    |                      |     |
                    |                      |     |
                    |                      |     |
   4           ###  +                      |     |
                    |                      |     |
                    |                      |     |
                ## T|                      |     |
                    |                      |     |
              .###  |                      |     |
   3          .###  +                      |     |
                    |                      |     |
              .###  |                      |     |
               ###  |                      |     |
             .####  |                      |     |
      .########### S|                      |     |
   2       .######  +                      |     |
             .####  |                      |     |
       .##########  |                      |     |
 L     .##########  |                      |-----|
 O     .##########  |                      |     |
 G    ############  |                      |     |
 I 1.############ X̄+T                      |     |
 T      .#########  |  I4                  |     |
       .##########  |  I22                 |  2  |
 M   .############  |S I11 I24             |     |
 E      .#########  |  I13 I16 I23 I3      |     |
 T        ########  |  I17 I5  I6          |     |
 R 0         #####  +X̄ I18 I20 I2  I7 I8   |     |
 I         ####### S|  I12 I15 I21         |-----|
 C         .######  |  I14 I19             |     |
            .#####  |S                     |     |
               ###  |                      |  1  |
                ##  |  I9                  |     |
  -1           .##  +T I10                 |     |
                .#  |  I1                  |     |
                .# T|                      |-----|
                 #  |                      |     |
                 #  |                      |     |
                    |                      |     |
  -2             .  +                      |     |
                 .  |                      |     |
                 .  |                      |     |
                    |                      |     |
                    |                      |     |
                    |                      |     |
  -3                +                      |     |
                    |                      |     |
                    |                      |     |
                    |                      |     |
                    |                      |     |
                 .  |                      |     |
  -4             .  +                      |  0  |
<LESS SELF-EFFICACY>   <EASY TO ENDORSE>
```

**Figure 4.** Variable map for a 4-point scale that was initially a 0 to 100 scale.
*Note*. # = 4 adolescents; . = 1 to 3 adolescents; $\bar{X}$ = mean; S = one standard deviation; T = two standard deviations; I = item; —— = threshold.

of logit: 1.23 ± 1.56), as demonstrated by the mismatch between person and item distributions. The results for the 0 to 100 scale are similar to those found for the 6-point scale and consequently not interpreted again.

### *Rating Scale Results for Pooled Sample*

Pooled results had person and item reliabilities of .95 and 1.00, with person and item separations of 4.48 and 14.69. Note, further collapsing of the 4-point scale resulted in a loss of both person and item separation. Average measures and thresholds were ordered, and outfit mean-square statistics ranged from 0.90 to 1.15. The 4-point rating scale structure is similar to the category probability curves in Figure 3. All infit and outfit item statistics fell within the recommended range of 0.5 to 1.5 ( $\bar{X}$ ± *SD* of Infit statistics = 1.01 ± 0.14, Minimum = 0.77, Maximum = 1.33; $\bar{X}$ ± *SD* of Outfit statistics = 0.99 ± 0.17, Minimum = 0.66, Maximum = 1.34).

The results are consistent with the variable maps (see Figures 2 and 4) provided for each sample. Because the pooled 4-point optimal scale results are similar to those obtained for each sample separately, the map is not interpreted again. In addition, a differential item functioning analysis as implemented in Winsteps was conducted and showed items behaved similarly for the two samples. A summary of the 24-item, 4-point scale item-level descriptive statistics is provided in Table 5, which shows most responses tended to be between the two middle categories and each item was descriptively symmetric.

## Discussion

The literature on self-efficacy response scales has led to inconsistent recommendations for measurement. This is partly due to contextual (e.g., domain, population) differences across studies and the particular measurement approach used (e.g., classical test theory, factor analysis, and generalizability theory). In this article, a Rasch modeling approach was used to evaluate the utility of two response scale formats used to measure middle school students' mathematics skills self-efficacy. What is the optimal number of categories? The results indicate that early adolescents' responses to two scales—a 100-point response scale recommended by Bandura (2006) in his Guide for Constructing Self-Efficacy Scales and a 6-point response scale frequently used in educational psychology research—are reduced to four categories. That is, the optimal number of response scale points after modifications were made was not equal to the number of rating scale points offered to early adolescents on the original printed version of the two self-efficacy forms,

**Table 5.** Descriptive Statistics for Items on the 24-item 4-Point Scale Using the Pooled Sample.

| | Category frequency | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item no. | 0 | 1 | 2 | 3 | $\bar{X}$ | *SD* | Skew | Kurtosis |
| 1 | 75 | 219 | 654 | 965 | 2.31 | 0.82 | −1.05 | 0.42 |
| 2 | 197 | 365 | 663 | 688 | 1.96 | 0.98 | −0.58 | −0.72 |
| 3 | 246 | 400 | 663 | 604 | 1.85 | 1.01 | −0.45 | −0.90 |
| 4 | 379 | 469 | 655 | 410 | 1.57 | 1.03 | −0.15 | −1.13 |
| 5 | 239 | 388 | 692 | 594 | 1.86 | 1.00 | −0.47 | −0.82 |
| 6 | 255 | 386 | 652 | 620 | 1.86 | 1.02 | −0.47 | −0.92 |
| 7 | 207 | 378 | 663 | 665 | 1.93 | 0.99 | −0.54 | −0.77 |
| 8 | 248 | 301 | 665 | 699 | 1.94 | 1.02 | −0.63 | −0.73 |
| 9 | 127 | 265 | 761 | 760 | 2.13 | 0.89 | −0.82 | −0.06 |
| 10 | 97 | 258 | 797 | 761 | 2.16 | 0.84 | −0.82 | 0.08 |
| 11 | 282 | 521 | 810 | 300 | 1.59 | 0.92 | −0.23 | −0.78 |
| 12 | 219 | 297 | 728 | 669 | 1.97 | 0.98 | −0.66 | −0.57 |
| 13 | 277 | 427 | 760 | 449 | 1.72 | 0.98 | −0.35 | −0.87 |
| 14 | 164 | 389 | 798 | 562 | 1.92 | 0.91 | −0.52 | −0.54 |
| 15 | 190 | 389 | 814 | 520 | 1.87 | 0.93 | −0.49 | −0.57 |
| 16 | 265 | 469 | 771 | 408 | 1.69 | 0.96 | −0.30 | −0.84 |
| 17 | 227 | 466 | 809 | 410 | 1.73 | 0.93 | −0.34 | −0.72 |
| 18 | 223 | 421 | 782 | 486 | 1.80 | 0.95 | −0.41 | −0.73 |
| 19 | 176 | 308 | 722 | 706 | 2.02 | 0.95 | −0.70 | −0.45 |
| 20 | 227 | 402 | 788 | 495 | 1.81 | 0.95 | −0.44 | −0.71 |
| 21 | 186 | 317 | 705 | 704 | 2.01 | 0.96 | −0.67 | −0.52 |
| 22 | 354 | 498 | 754 | 306 | 1.53 | 0.97 | −0.17 | −0.96 |
| 23 | 308 | 468 | 803 | 332 | 1.61 | 0.95 | −0.26 | −0.85 |
| 24 | 360 | 490 | 691 | 370 | 1.56 | 1.01 | −0.15 | −1.05 |

suggesting that more than four categories is not optimal for this age group. Our findings are consistent with those reported by E. V. Smith et al. (2003) who found that upper elementary students made use of four categories despite being given a 0 to 100 response scale in 10-unit increments to assess their writing self-efficacy. Moreover, the fact that self-efficacy data from two independent samples reduced to four categories provides preliminary evidence of stability and generalizability.

One possible explanation for this finding is that, when students judge what they can do, their judgment boils down to four basic categories: I *cannot* do this, I'm *not sure that I can* do this, I am *pretty sure I can* do this, I *can* *definitely* do this. Perhaps children's efficacy judgments are simply not more nuanced than that. Another explanation is that early adolescents' working memory capacity is limited to three to five categories, and offering any more response options induces unnecessary cognitive burden (Cowan, 2010).

As noted earlier, it could be that the greater one's level of expertise, the greater one's ability to make more fine-tuned self-judgments. Older students (e.g., undergraduates and graduates) who are more familiar with the demands of a given academic domain such as mathematics might be better equipped to make more nuanced judgments about their capabilities to handle those demands (Weil et al., 2013). On the other hand, domain expertise is not the only factor that explains whether individuals think of their efficacy in more complex terms. Individuals who possess a heightened level of awareness of themselves and their capacities may be able to make more nuanced evaluations of what they can do. Researchers could investigate expertise both in terms of the domain in question and in terms of a learner's self-knowledge. The latter may be assessed by measuring metacognitive awareness, which could in part account for individual variation in response style (Schraw, 2009).

Readers might ask whether there is a practical advantage to a 4-point scale over a 6-point scale. We believe there is a benefit. Most obviously, providing early adolescents with fewer response categories lessens their cognitive burden and thereby increases the likelihood that they will complete surveys. Furthermore, our findings suggest that researchers may not gain, and may indeed lose, in their understanding of self-efficacy and its correlates when they use response scales that contain too many categories. Including more than four categories or too many categories on a self-efficacy response scale might lead to unsystematic measurement error or less information (E. V. Smith et al., 2003) and possibly correlations and effect sizes that may be misleading. Given the findings of Embretson (1996) and Kang and Waller (2005), we might speculate that results would be misleading in the sense of inflated correlations and effect sizes. Based on our findings, we recommend that researchers studying self-efficacy and related motivation constructs with early adolescents use the approach used in this article and by E. V. Smith et al. (2003) for optimizing the number of rating categories of instruments. The benefits of using this approach are increased reliability and validity about group and individual level inferences.

A secondary finding to emerge from our study was that categories in the middle of the response scale tended to collapse more often than those toward the anchor labels. This suggests that labeling each response option, and not only endpoints, might help avoid ambiguity in the meaning of response categories. Leaving categories unlabeled may introduce error, "as the meanings of these unlabeled intermediate categories must be created by each individual, which may lead to undesirable response sets" (E. V. Smith et al., 2003, p. 387). Moreover, findings suggest that although fewer categories may increase the validity of early adolescents' responses, collapsing too many categories results in a loss of separation and reliability and is not optimal. Also, too many categories can introduce avoidable measurement error (E. V. Smith et al., 2003). The collapsing process used in this article is in fact a means of showing how and where this unreliability arises in data. Based on our findings, we recommend that early adolescence measures be routinely checked for this sort of loss of information, particularly if category collapsing is performed.

Social scientists who are accustomed to using wider-ranging response scales to measure psychological phenomena may wonder whether a similar response pattern would emerge in their data. We agree that this is a possibility worth examining. Still, our results must be considered in light of our multifaceted study context. First, we assessed self-efficacy, which may be germane to four basic levels of cognitive awareness about one's capabilities. However, as noted in the introduction, self-efficacy can be operationalized in more specific or more general terms. Here self-efficacy was assessed in terms of perceived capability to solve problems related to a set of mathematics topic. Mathematics self-efficacy can also be evaluated at more general or specific levels. Students can judge their efficacy for doing well in mathematics generally, in specific mathematics courses, or in terms of learning mathematics. Alternatively, they can rate their sense of efficacy for doing specific mathematics problems. When self-efficacy is assessed at other levels of specificity, greater or fewer response categories may be more appropriate.

As did E. V. Smith et al. (2003) who reported similar findings in the area of writing, we examined self-efficacy in the context of early adolescence and with predominantly White samples. Older students (i.e., undergraduate and graduate students) and students of different backgrounds (e.g., cultural, geographical, ethnic) may respond differently. The academic domain in question may also influence students' use of response categories. Moreover, the observation that at least half of the early adolescents in our study tended to be above the locations of most items on the self-efficacy continuum (see Figures 2 and 4) could possibly explain why not all response categories are being used as expected and were therefore collapsed to four categories. Conducting a similar study of self-efficacy among low-achieving students with or without a reduced rating scale system might render different results.

The method used for combining categories in the 0- to 100-point scale started with empirically flagging disordered categories and then collapsing them using substantive reasoning (i.e., combining categories that were viewed to signify the same level of the trait by the researchers), but findings might still be sample dependent because of the strong dependency on flagging category order with empirical results. Moreover, the approach we used for collapsing categories was mostly descriptive and is a limitation of the current study. Altering any one of the contextual variables mentioned above (and indeed, other contextual variables that were not considered) might result in differing response patterns. Similarly, the results we obtained may depend on the fact that we started our analyses with more than four categories (E. V. Smith et al., 2003). Beginning with four categories might not necessarily lead to the same result as starting with a 1 to 6 or 0 to 100 scale and collapsing the responses to four categories. This conclusion needs additional verification in other contexts.

The generalizability of the findings from this study is limited by the mismatch between the distribution of items compared with early adolescents' self-efficacy levels (shown in Figures 2 and 4). This mismatch raises some concern about the ability of this instrument to assess self-efficacy precisely, especially for those with higher levels of self-efficacy. A 4-point response scale solution may only be relevant in similar cases where items are easy to endorse. This is not an indication that our new instrument is not useful, but that inferences regarding higher self-efficacy should be made with some caution given the lack of precision at this level, particularly if summed (raw) scores are used (see Kang & Waller, 2005). Overall, this finding is not unique to our study and is a known concern in published studies of self-efficacy in other domains (e.g., Alviar-Martin, Randall, Usher, & Engelhard, 2008; E. V. Smith et al., 2003). Therefore, we recommend that future studies use a 4-point scale that includes more challenging-to-endorse items to better match the range of self-efficacy levels being assessed and to improve the generalizability of findings.

Given that the 0- to 100-point scale is popular in self-efficacy research and often assumed continuous by most applied researchers, it would be an interesting endeavor for researchers to consider using another model such as Müller's (1987) Rasch model for continuous ratings. Specifically, researchers could examine the utility of this model with 0- to 100-point self-efficacy scales in various domains (e.g., reading, writing) and investigate how the scale compares to those based on discrete polytomous models such as the rating scale model. Addressing these questions was beyond the scope of the current study.

We are certain that findings from continued analyses of this type will benefit researchers as they choose the number of rating categories appropriate for assessing mathematics skills self-efficacy among middle school students. Findings from this study may also be applicable to those who study related motivation constructs. Using response scales with fewer response categories will have implications for the inferential statistical procedures used by social scientists. Although researchers typically treat Likert-type response scales as continuous, this assumption, particularly when made in conjunction with shorter (i.e., 4-point) response scales, is faulty. The wiser practice is to use Rasch or IRT techniques to calibrate items and score item responses based on a Likert-type scale. Most researchers in the social sciences may not feel comfortable scaling item response using Rasch or IRT techniques, but some helpful resources for researchers wanting to learn about how to optimize or evaluate their rating scale using Rasch techniques are provided by E. V. Smith et al. (2003). Researchers wanting a practical guide to conducting IRT analyses may refer to Toland (2014). For more in-depth information about IRT and Rasch, researchers can refer to de Ayala (2009). Whatever their analytical choice, researchers should keep in mind that the inferences drawn about a construct such as self-efficacy are only as good as the measurement instrument used to reflect that construct. Therefore, we recommend that researchers provide empirical evidence for the response scale used and thereby improve statistical conclusion and construct validity.

## Appendix A

**Directions:** Using the same scale, please rate how much **confidence you have that you can succeed at exercises related to the following math topics** *without using a calculator*. Remember that *you can circle any number from* 1 (*not confident at all)* to 6 (*completely* confident).

1 2 3 4 5 6

Not at all confident Completely confident

| | How confident are you that you can successfully solve math exercises involving . . . | Not at all confident | | | Completely confident | | |
|---|---|---|---|---|---|---|---|
| 1 | Multiplication and division | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | Decimals | 1 | 2 | 3 | 4 | 5 | 6 |
| 3 | Fractions | 1 | 2 | 3 | 4 | 5 | 6 |
| 4 | Ratios and proportions | 1 | 2 | 3 | 4 | 5 | 6 |
| 5 | Percents | 1 | 2 | 3 | 4 | 5 | 6 |
| 6 | Powers and exponents | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | Factors and multiples | 1 | 2 | 3 | 4 | 5 | 6 |
| 8 | Inequalities (>, <, ≤, ≥, ≠) | 1 | 2 | 3 | 4 | 5 | 6 |

*(continued)*

## Appendix A (continued)

| How confident are you that you can successfully solve math exercises involving . . . | | Not at all confident | | | | | Completely confident |
|---|---|---|---|---|---|---|---|
| 9 | Order of operations | 1 | 2 | 3 | 4 | 5 | 6 |
| 10 | Rounding and estimating | 1 | 2 | 3 | 4 | 5 | 6 |
| 11 | Word problems | 1 | 2 | 3 | 4 | 5 | 6 |
| 12 | Equations with one variable | 1 | 2 | 3 | 4 | 5 | 6 |
| 13 | Equations with two or more variables | 1 | 2 | 3 | 4 | 5 | 6 |
| 14 | Graphing | 1 | 2 | 3 | 4 | 5 | 6 |
| 15 | Tables, charts, diagrams, and coordinate grids | 1 | 2 | 3 | 4 | 5 | 6 |
| 16 | Angles, perimeter, area, and volume | 1 | 2 | 3 | 4 | 5 | 6 |
| 17 | Multi-step problems | 1 | 2 | 3 | 4 | 5 | 6 |
| 18 | Measurement | 1 | 2 | 3 | 4 | 5 | 6 |
| 19 | Mean, median, range, and mode | 1 | 2 | 3 | 4 | 5 | 6 |
| 20 | Chance and probability | 1 | 2 | 3 | 4 | 5 | 6 |
| 21 | Negative numbers | 1 | 2 | 3 | 4 | 5 | 6 |
| 22 | Explaining in words how you solved a math problem | 1 | 2 | 3 | 4 | 5 | 6 |
| 23 | Using math in other subjects | 1 | 2 | 3 | 4 | 5 | 6 |
| 24 | Doing quick calculations in your head | 1 | 2 | 3 | 4 | 5 | 6 |

## Appendix B

**Directions:** On a scale from **0** (***not at all confident***) to **100** (***completely confident***), please rate how much **confidence you have that you can succeed at exercises related to the following math topics** *without using a calculator*. Please *write in* any number between 0 and 100.

| 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Not at all confident | | | | | Somewhat confident | | | | | Completely confident |

| | How confident are you that you can successfully solve math exercises involving . . . | Confidence rating (0-100) |
|---|---|---|
| 1 | Multiplication and division | |
| 2 | Decimals | |
| 3 | Fractions | |

*(continued)*

## Appendix B (continued)

| | How confident are you that you can successfully solve math exercises involving . . . | Confidence rating (0-100) |
|---|---|---|
| 4 | Ratios and proportions | |
| 5 | Percents | |
| 6 | powers and exponents | |
| 7 | Factors and multiples | |
| 8 | Inequalities (>, <, ≤, ≥, ≠) | |
| 9 | Order of operations | |
| 10 | Rounding and estimating | |
| 11 | Word problems | |
| 12 | Equations with one variable | |
| 13 | Equations with two or more variables | |
| 14 | Graphing | |
| 15 | Tables, charts, diagrams, and coordinate grids | |
| 16 | Angles, perimeter, area, and volume | |
| 17 | Multi-step problems | |
| 18 | Measurement | |
| 19 | Mean, median, range, and mode | |
| 20 | Chance and probability | |
| 21 | Negative numbers | |
| 22 | Explaining in words how you solved a math problem | |
| 23 | Using math in other subjects | |
| 24 | Doing quick calculations in your head | |

## 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

Alviar-Martin, T., Randall, J. D., Usher, E. L., & Engelhard, G. (2008). Teaching civic topics in four societies: Examining national context and teacher confidence. *The Journal of Educational Research*, *101*, 177-188. doi:10.3200/JOER.101.3.177-188

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. Retrieved from http://link.springer.com/journal/11336

Andrich, D. (2013). An expanded deviation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy." *Educational and Psychological Measurement*, *73*, 78-124. doi:10.1177/0013164412450877

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Adolescence and education, Vol. 5: Self-efficacy and adolescence* (pp. 307-337). Greenwich, CT: Information Age.

Bong, M. (2001). Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary Educational Psychology*, *26*, 553-570. doi:10.1006/ceps.2000.1048

Bong, M. (2006). Asking the right question: How confident are you that you could successfully perform these tasks? In F. Pajares & T. Urdan (Eds.), *Adolescence and education, Vol. 5: Self-efficacy beliefs of adolescents* (pp. 287-305). Greenwich, CT: Information Age.

Bong, M., & Hocevar, D. (2002). Measuring self-efficacy: Multitrait-multimethod comparison of scaling procedures. *Applied Measurement In Education*, *15*, 143-171. doi:10.1207/s15324818AME1502_02

Brown, S. D., & Lent, R. W. (2006). Preparing adolescents to make career decisions: A social cognitive perspective. In F. Pajares & T. Urdan (Eds.), *Adolescence and education, Vol. 5: Self-efficacy beliefs of adolescents* (pp. 201-223). Greenwich, CT: Information Age.

Butz, A. R., Toland, M. D., Zumbrunn, S. K., Danner, F. W., & Usher, E. L. (2014, April). *What is the "magic number?" A review of response categories in measuring writing self-efficacy.* Paper presented at the meeting of the American Educational Research Association, Philadelphia, PA.

Cheema, J. R., & Kitsantas, A. (2013). Influences of disciplinary classroom climate on high school student self-efficacy and mathematics achievement: A look at gender and racial-ethnic differences. *International Journal of Science and Mathematics Education*, *12*, 1261-1279.

Cipriani, D. J., Hensen, F. E., McPeck, D. L., Kubec, G. L. D., & Thomas, J. J. (2012). Rating scale analysis and psychometric properties of the caregiver self-efficacy scale for transfers. *Physical & Occupational Therapy in Pediatrics*, *32*, 405-415. doi:10.3109/01942638.2012.694993

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51-57. doi:10.1177/0963721409359277

Cowan, N., Morey, C. C., Chen, Z., Gilchrist, A. L., & Saults, J. S. (2008). Theory and measurement of working memory capacity limits. *The Psychology of Learning and Motivation*, *49*, 49-104. doi:10.1016/S0079-7421(08)00002-9 de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: SAGE.

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, *53*, 414-439.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*, 201-212. doi:10.1177/014662169602000302

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, *20*, 261-273. Retrieved from http://www.nctm.org/publications/toc.aspx?jrnl=jrme

Hagquist, C., Bruce, M., & Gustavsson, J. P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, *46*, 380-393. doi:10.1016/j.ijnurstu.2008.10.007

Kan, A. (2009). Effect of scale response format on psychometric properties in teaching self-efficacy. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, *34*, 215-228. Retrieved from http://www.ejer.com.tr/index.php

Kang, S.-M., & Waller, N. G. (2005). Moderate multiple regression, spurious interaction effects and IRT. *Applied Psychological Measurement*, *29*, 87-105. doi:10.1177/0146621604272737

Klassen, R. M., & Usher, E. L. (2010). Self-efficacy in educational settings: Recent research and emerging directions. In T. C. Urdan & S. A. Karabenick (Eds.), *ADVANCES IN MOTIVATION AND ACHIEVEMENT: Vol. 16A. The decade ahead: Theoretical perspectives on motivation and achievement* (pp. 1-33). Bingley, UK: Emerald.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106. Available from http://www.jampress.org/

Linacre, J. M. (2003). Data variance: Explained, modeled, and empirical. *Rasch Measurement Transactions*, *17*, 942-943. Retrieved from http://www.rasch.org/rmt/rmt173g.htm

Linacre, J. M. (2009). *A user's guide to Winsteps, Ministep, Rasch-model computer programs: Program manual 3.72.3*. Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Linacre, J. M. (2011). Winsteps (Version 3.72.0) [Computer Software]. Beaverton, OR: Winsteps.com. Available from http://www.winsteps.com/

Lozano, L. M., Garciá-Cueto, E., & Muñiz, J. (2008). Effect of number of response categories on the reliability and validity of rating scales. *Methodology*, *4*, 73-79. doi:10.1027/1614-2241.4.2.73

Maurer, T. J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology*, *83*, 324-329. doi:10.1037/0021-9010.83.2.324

Morony, S., Kleitman, S., Lee, Y. P., & Stankov, L. (2013). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research*, *58*, 79-96.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165-181.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics: E-standards*. Available from http://standards.nctm.org

Pajares, F., & Barich, J. (2005). Assessing self-efficacy: Are skills-specific measures better than domain specific measures? *Psychology*, *12*, 334-348.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, *24*, 124-139. doi:10.1006/ceps.1998.0991

Pajares, F., Hartley, J., & Valiante, G. (2001). Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development*, *33*, 214-221. Available from http://mec.sagepub.com/

Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics outcomes: The need for specificity of assessment. *Journal of Counseling Psychology*, *42*, 190-198. doi:10.1037/0022-0167.42.2.190

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33-40. doi:10.1037/0022-0663.82.1.33

Schneider, W. (2008). Children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*, 114-121. doi:10.1111/j.1751-228X.2008.00041.x

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33-45.

Schunk, D. H., & Pajares, F. (2005). Competence beliefs in academic functioning. In A. J. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 85-104). New York, NY: Guilford Press.

Smith, E. V., Jr., Wakely, M. B., De Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement*, *63*, 369-391. doi:10.1177/0013164403251320

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*, 516-517. Retrieved from http://www.rasch.org/rmt/

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*, 199-218. Available from http://www.jampress.org/

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*, 120-151. doi:10.1177/0272431613511332

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, *34*, 89-101. doi:10.1016/j.cedpsych.2008.09.002

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., . . .Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*, 264-271.

Wright, B. D., & Linacre, J. M. (1992). Combining and splitting categories. *Rasch Measurement Transactions*, *6*, 233-235. Retrieved from http://www.rasch.org/rmt/contents.htm

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

## Author Biographies

**Michael D. Toland** is an associate professor in the Educational Psychology program in the Department of Educational, School, and Counseling Psychology at the University of Kentucky. His research interests include psychometrics, item response theory, factor analysis, scale development, multilevel modeling, and the realization of modern measurement and statistical methods in educational research.

**Ellen L. Usher** is an associate professor in the Educational Psychology program in the Department of Educational, School, and Counseling Psychology at the University of Kentucky and director of the P20 Motivation and Learning Lab. Her research focuses on the sources and effects of beliefs of personal efficacy from the perspective of social-cognitive theory.