Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 7, May 2018 ISSN 1531-7714

An Application of the Partial Credit IRT Model in Identifying Benchmarks for Polytomous Rating Scale Instruments

Enis Dogan, *National Center for Education Statistics*

Several large scale assessments include student, teacher, and school background questionnaires. Results from such questionnaires can be reported for each item separately, or as indices based on aggregation of multiple items into a scale. Interpreting scale scores is not always an easy task though. In disseminating results of achievement tests, one solution to this conundrum is to identify cut scores on the reporting scale in order to divide it into achievement levels that correspond to distinct knowledge and skill profiles. This allows for the reporting of the percentage of students at each achievement level in addition to average scale scores. Dividing a scale into meaningful segments can, and perhaps should, be done to enrich interpretability of scales based on questionnaire items as well. This article illustrates an approach based on an application of Item Response Theory (IRT) to accomplish this. The application is demonstrated with a polytomous rating scale instrument designed to measure students' sense of school belonging.

In addition to cognitive items that are aimed at measuring student achievement in subjects such as Reading and Mathematics, several large scale assessments (e.g. National Assessment of Educational Progress, Trends in International Mathematics and Science Study) also include what are known as background questionnaires. These are usually polytomous rating scale instruments that ask students to indicate their degree of affirmation with several statements related to the target construct. These instruments provide "additional information that helps put student achievement results into context and allows meaningful comparison between student groups" (National Center for Education Statistics, n.d.). Results from such instruments can be reported for each item separately, or as indices based on aggregation of multiple items into a scale. Interpreting scale scores is not always an easy task though: what does a score of 6.7 on a scale of 0 to 10 mean exactly? In disseminating results of achievement tests, one solution to this conundrum is to identify cut scores on the reporting scale in order to divide it into achievement levels that correspond to distinct knowledge and skill profiles. This allows for the reporting of the percentage of students at each achievement level in addition to average scale scores such as Basic, Proficient and Advanced.

Dividing a scale into meaningful segments and reporting the percentage of students/respondents in each can, and perhaps should, be done to enrich interpretability of scales based on rating scale instruments as well. This article illustrates an application of Item Response Theory (IRT) in identifying benchmarks (i.e., cut scores) on such scales in order to divide the scale into meaningful and interpretable segments. The proposed approach can be applied if the items that make up the scale of interest are calibrated with a proper polytomous IRT model to a common metric, including the Partial Credit Model (PCM; Masters and Wright, 1997), Graded Response Model (GRM; Samejima, 1969), and Rating Scale Model (RSM; Andrich, 1978) among others. Evidently, items must all be measuring the same underlying construct and the assumptions of the IRT model of choice must be met before the proposed approach can be implemented.

Applications of IRT in rating scale instruments

Although more commonly used with achievement tests, IRT models can also be used for both item calibration and "ability" estimation with rating scale instruments such as questionnaires measuring psychological or behavioral constructs[1]. There are several advantages to using IRT models with rating scale instruments compared to other approaches such as sum scoring. Most importantly, such instruments yield data that are ordinal, not interval, making use of sum scores questionable (Smith, Conrad, Chang, & Piazza, 2002). In addition, IRT models provide "sample-free measurement estimates, making it possible to estimate a person's level of the latent construct free of the distribution of the individual items and to estimate an item's difficulty level free from the distribution of people used in the sample" (DiStefano and Morgan, 2011, p.356). In addition, as Reeve and Masse (2004) point out, these models allow more in-depth analysis of items, examination of precision across the score continuum (as opposed to a single overall reliability coefficient) and better handling of complex measurement problems such as linking scores across alternative forms.

There are numerous studies in the literature that utilized IRT modeling with rating scale instruments. For example, Amtmann et al. (2010) evaluated the psychometric properties of the Patient-Reported Outcomes Measurement Information System Pain Interference (PROMIS-PI) item bank using the GRM. Based on analyses of dimensionality, item fit, differential item functioning, scale information functions (precision), associations between PROMIS-PI scores and other measures, the researchers concluded that the PROMIS-PI items constitute a psychometrically sound item bank for assessing the negative effects of pain.

A more recent example of the use of IRT with rating scale instruments come from Anthony, DiPerna, and Lei (2016). They applied the GRM to the Social Skills Improvement System — Teacher Rating Scale, a measure of student social skills. They used IRT-based item analysis, item and test information functions and item fit statistics to select a subset of items that yielded equivalent reliability and validity evidence compared to the published version of the scale, which can be completed in approximately half the time. Other similar applications include Jong et al. (2015), who used the RSM in scaling an instrument measuring preservice teachers' dispositions, attitudes, and beliefs about mathematics teaching and learning, Bonanomi et al. (2018), who used a multidimensional RSM and multidimensional PCM to investigate the construct validity of an instrument measuring high school students' learning motivation, and a study by Carmichael et al. (2010) that examined psychometric properties of an instrument assessing middle school students' interest in statistical literacy using the RSM.

Identifying cut scores for rating scale instruments

As mentioned earlier, this study aims to establish benchmarks on scales derived from rating scale instruments in order to divide the scale into meaningful and interpretable segments. There are earlier studies that also identified the need for meaningful benchmarks in interpreting scores from such instruments. For instance, in discussing the need for meaningful ways of interpreting scores based on patient-reported outcomes (PROs), Morgan et al. (2017) noted that "As PROs move from the realm of clinical research and clinical trials to use in patient care, a framework for score interpretation is required" (p. 566). They further argued that "Despite their strong psychometric properties, the lack of an empirically established framework to interpret PROMIS scores in a clinically meaningful way impedes their use" (p.566). These researchers applied the modified bookmark standard setting method, a well-established approach for standard setting in educational testing using IRT-calibrated items, to identify cut-points for PROMIS pediatric measures of physical health. The bookmark method relies on Ordered Item Booklets that contain items ordered by difficulty from easiest to the most difficult. The method requires subjective judgment of matter experts, where a panel of experts are asked to place a bookmark between two items such that the "minimally qualified" respondent for a particular category is expected to endorse the items below the bookmark and not to endorse the items above the bookmark (Karantonis & Sireci, 2006). Results of the Morgan et al. (2017) study were unfortunately mixed. The cut-scores were not consistent among panels of parents, patients and clinicians for some of the measures, likely due to the subjective nature of the method used. Another study that used the bookmark

[1]Ability in this case refers to the location of a given individual on the underlying scale of interest.

method is from Cook et al. (2015), who identified cut scores for a PRO measure of fatigue, physical functioning, and sleep disturbance. They found that patient and clinician panels set identical cut scores for severity levels of lower extremity function and sleep disturbance, but their cut scores were 0.5 SD apart for upper extremity function and fatigue.

DiStefano and Morgan (2011) compared three different methods of creating cut scores for a teacher-reported measure of student behavioral and emotional problems. They used T scores, receiver operating characteristic curve analysis, and the RSM in identifying cut scores. With the RSM, scores in the 60th to 90th percentiles were examined as potential cutoff scores. The score corresponding to the 65th percentile provided the highest levels of sensitivity and specificity. Researchers found that the three methods were generally in concordance. In a similar study, Yovanoff and Squires (2006) compared ROC and Rasch methods of creating cut scores on a social-emotional screening test. They too found that the two approaches yielded similar results.

The methods used in studies above have limitations in the context of rating scale instruments. Bookmark and similar methods rely on expert judgment. These methods are more commonly used in setting cuts scores for achievement tests where there is always a correct answer to each item. This makes it relatively easier for the experts to decide whether a student can correctly (or partially correctly) answer a given item based on the description of the student's knowledge and skills (as reflected in Achievement Level Descriptors) and the knowledge and skills required to successfully answer that item. With rating scale items, it is harder to make such judgments. This results in inconsistencies in identifying cut scores as evidenced by Morgan et al. (2017) and Cook et al. (2015). On the other hand, methods based on normative criteria, exemplified in DiStefano and Morgan (2011) lack criterion-based interpretation. The approach presented in this study yields criterion-based cut scores without the need for subjective expert judgement.

Method

The item characteristic curve (ICC) is the basis of IRT. An ICC is a logistic function that displays the probability that a respondent will endorse a particular response option (e.g., Strongly Agree) given his/her "ability" and item characteristics, such as difficulty and discrimination. Figure 1 below shows an ICC for a hypothetical item with four answer options, calibrated

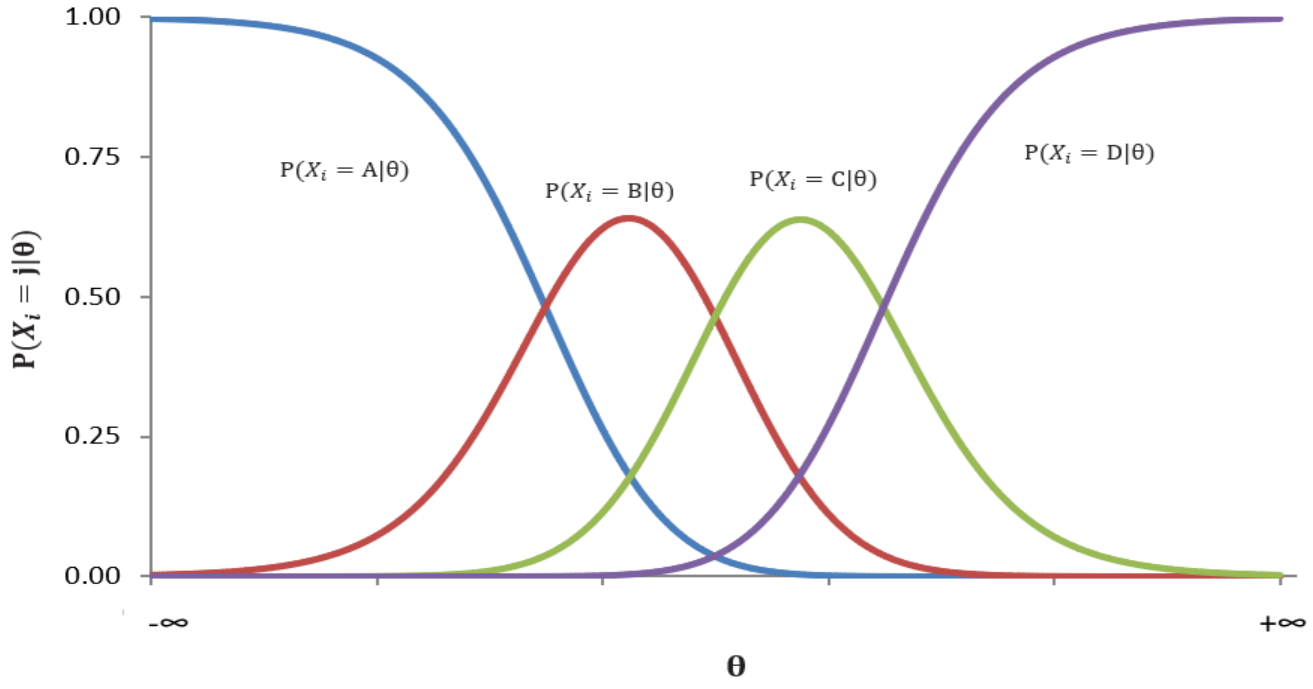

Figure 1: Item Characteristic Curves for a hypothetical item with four answer options, calibrated with the Partial Credit Model.

with the Partial Credit Model. The y-axis denotes $P(X_i=j|\theta)$, the probability that a respondent with ability of θ would endorse answer option *j* in response to item *i*:

$$P(X_i = j|\theta) = e^{\sum_{j=0}^{m}(\theta - \beta_i + \tau_{ij})} / \sum_{h=0}^{m} e^{\sum_{j=0}^{m}(\theta - \beta_i + \tau_{ij})},$$

where the item parameter β_i gives the location of item i on the latent construct and τij denotes step parameters for the response levels, ranging $j = 0$ to $j = m$ for the same item. Note that in the partial credit model all discrimination parameters (a) are set to 1. At extremely low values of θ, the expected response is almost certainly A. Similarly, at extremely high values of θ, the expected response is almost certainly D. For a certain range in the middle of the scale, responses B and C become the most likely response. Note that any given point on the scale,

$$\sum_{j=0}^{m} P(X_i = j|\theta) = 1.$$

Once the ICCs for each response option for each item are plotted, they can be summed across items to create a Scale Characteristics Curves (SCCs). A hypothetical SCC for a scale made up of n items, each with four response options, A (most negative) to D (most positive), is displayed in Figure 2. The x-axis of an SCC still represents the θ scale; however, the y-axis is no longer a probability. For each curve, the y-axis represents the number of items expected to be endorsed as the given response option conditional on θ:

$$E(j|\theta) = \sum_{i=1}^{i=n} P(X_i = j|\theta)$$

where *i* represents the items, ranging from 1 to *n*, and j represents the response options, ranging from 1 to m. Therefore at any given point on θ,

$$\sum_{i=1}^{i=n} E(j|\theta) = \sum_{j=0}^{m} \sum_{i=1}^{i=n} P(X_i = j|\theta) = n.$$

Hence the, the maximum value on y-axis in Figure 2 is equal to n.

Once the SCCs are plotted, the next step is to identify the points where the curves intersect such that $P(X_i = j) = P(X_i = j + 1)$. These points divide the θ scale into *m*-1 segments. In Figure 2, the four curves intersect at three points and divide the scale into four segments: (1) $\theta < \theta_1$, (2) $\theta_1 \leq \theta < \theta_2$, (3) $\theta_2 \leq \theta < \theta_3$, and (3) $\theta \geq \theta_3$. These points of intersection can be used as

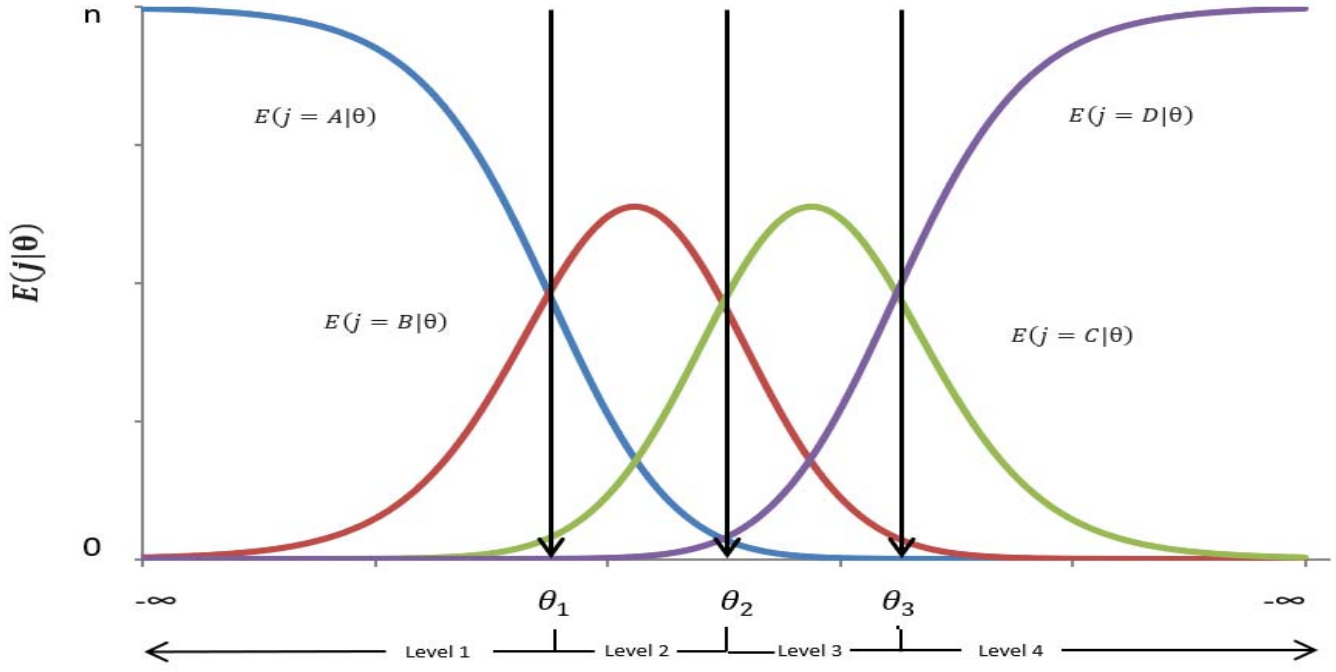

Figure 2: A hypothetical Scale Characteristics Curves for a scale made up of n items, each with four answer options.

benchmarks (i.e., cut scores) to divide the scale into four levels that can be described as follows:

- **Level 4 (Most Positive):** Across the n items, students at this level are more likely to endorse [D] than any other single response option
- **Level 3 (Positive):** Across the n items, students at this level are more likely to endorse [C] than any other single response option
- **Level 2 (Negative):** Across the n items, students at this level are more likely to endorse [B] than any other single response option
- **Level 1 (Most Negative)**: Across the n items, students at this level are more likely to endorse [A] than any other single response option

The approach is illustrated below using published item parameter estimates from the 2015 TIMSS student background questionnaire for the Students' Sense of School Belonging (SSSB) Scale.

Instruments

It is crucial to note that the approach introduced in this study requires that a reliable and valid scale has been established in advance. In establishing such scales with IRT modeling, unidimensionality and monotonicity of the scale, model and item fit must be investigated first (Bond & Fox, 2007). Below is a discussion of analyses of the SSSB Scale, all conducted by TIMSS, as reported in Martin et al. (n.d.).

The SSSB Scale is intended to reflect students' feelings towards their school and connectedness with the school community. Students participating the TIMSS assessments in 2015 (n=300,000) were asked to indicate the degree of their agreement with each of the seven statements (Table 1) that make up the scale: Agree a lot, Agree a little, Disagree a little, or Disagree a lot. TIMSS constructed the SSSB scale using the PCM. Although the complete student background questionnaire features a larger set of items, in constructing the SSSB scale, the seven items that made up the scale were calibrated on their own, in absence of the other questionnaire items. This was necessary since other items in the questionnaire measured constructs different from sense of school belonging. Cronbach's Alpha for the scale was .82 for the US sample. The inter-item correlations ranged from .27 to .65 with a median of .42. The corrected item to total score correlations ranged from .47 (like to see classmates) to .72 (proud to go to this school). In investigating unidimensionality, TIMSS conducted a Principal Component Analysis. The first principal component had an eigenvalue of 3.63. The eigenvalue for the second component was 0.83. The component loadings for the first component ranged from .57 (like to see classmates) to .82 (proud to go to this school). The SSSB scale was centered at 10, the mean score across all TIMSS countries. The standard deviation of the scale was set to 2. This was achieved through a linear transformation of the logit scale score, where Transformed Scale Score = 7.847376 + 1.363355 * Logit Scale. Table 1 below displays the parameter estimates for the seven items that make up this scale along with Rasch infit item statistic, a mean-square residual summary statistics indicating item misfit. Infit item statistics ranged from 0.91 to 1.17, satisfying the

Table 1: Item parameter estimates for the Students' Sense of School Belonging Scale: Grade 8 TIMSS 2015 assessment.

Item ID	What do you think about your school? Tell how much you agree with these statements.	b	d1	d2	d3	infit
BSBG15A	I like being in school	0.38	-0.96	-0.74	1.70	1.01
BSBG15B	I feel safe when I am at school	0.07	-0.95	-0.59	1.53	0.99
BSBG15C	I feel like I belong at this school	0.21	-0.84	-0.55	1.39	0.94
BSBG15D	I like to see my classmates at school	-0.73	-0.52	-0.47	0.99	1.17
BSBG15E	Teachers at my school are fair to me	0.20	-0.98	-0.56	1.54	1.12
BSBG15F	I am proud to go to this school	0.27	-0.76	-0.50	1.27	0.91
BSBG15G	I learn a lot in school	-0.40	-0.90	-0.62	1.52	0.98

Source: Martin et al. (n.d.).

b represents the estimate for β and $d_j$represent the estimate for the τ_j parameter for the given item.

criteria offered by Bond & Fox (2007) that sets values between 0.7 and 1.3 as acceptable. In addition, the SSSB scale scores were positively correlated with both TIMSS mathematics and science scores at r=.14 and r=.13, respectively, providing external validity evidence.

Results

In applying the approach described above, ICCs for each response option of each item were plotted and summed across the seven items to create a Scale Characteristics Curves (SCCs) for this scale (Figure 3). E(j|θ) curves intersected at following points:

- E(j=Agree a lot|θ) and E(j=Agree a little|θ) intersected at θ = 1.42, corresponding to a scale score of 9.78.
- E(j=Agree a little|θ) and E(j=Disagree a little|θ) intersected at θ = -.72, corresponding to scale scores of 6.85
- E(j=Disagree a little|θ) and E(j=Disagree a lot|θ) intersected at θ = -.73, corresponding to scale scores of 6.86.

Since the last two intersection points are nearly identical, SSSB scale scores of 9.78 and 6.86 were used as benchmarks to divide the scale into three levels. The distance between the two cut scores is 2.92 scale score points, which corresponds to 1.46 standard deviations on the scale. Level 1 corresponds to all points on the scale below 6.86. Level 2 corresponds to all points greater than or equal to 6.86 and smaller than 9.78. Level 3 corresponds to all points on the scale greater than or equal to 9.78.

These three levels that can be defined as follows:

- Level 3 (Most Positive): Across the seven items, students at this level are more likely to endorse *Agree a lot* than any other single response option
- Level 2 (Positive): Across the seven items, students at this level are more likely to

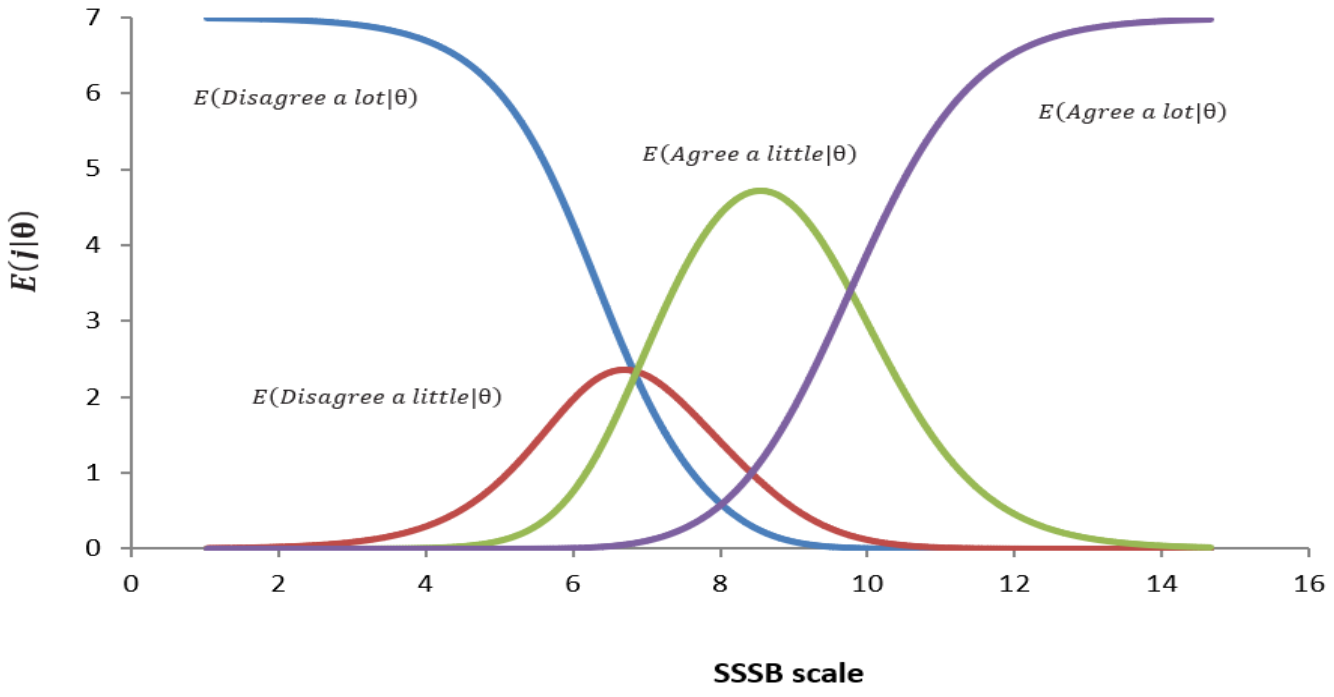

Figure 3: Scale Characteristic Curves for the Students' Sense of School Belonging Scale: Grade 8 TIMSS 2015 assessment.

endorse *Agree a little* than any other single response option

- Level 1 (Negative): Across the seven items, students at this level are more likely to endorse either *Disagree a little* or *Disagree a lot* than any other single response option

Based on the cut scores identified above, 6.1% of the students in the TIMSS 2015 US sample were classified to Level 1, 47.2% were classified to level 2 and 46.6% were classified to Level 3. Table 2 displays the observed distribution of response options for each item by trait level. At Level 3, the most frequently endorsed answer was *Agree a lot* for six of the seven items. The exception was item 1 (I like being in school), where the most frequent answer was *Agree a little* (50%), followed by *Agree a lot* (%45). At Level 2, the most frequently endorsed answer was *Agree a little* for six of the seven items. The exception was item 4 (I like to see my classmates at school l), where the most frequent answer was *Agree a lot* (45%), followed by *Agree a little* (%41). At Level 1, the most frequently endorsed answer was *Disagree a lot* for six of the seven items. The exception was item 4 (I like to see my classmates at school l), where the most frequent answer was *Agree a little* (32%), followed by *Disagree a lot* (%31). These results provide validity evidence for the description of the trait levels.

Comparisons to TIMSS benchmarks

As mentioned above, TIMSS also divides the SSSB scale into three levels: Little Sense of School Belonging, Sense of School Belonging, and High Sense of School Belonging. First an expected total score was defined as follows:

$$\mathrm{E(Total|\theta)} = \sum_{j=0}^{m} j \left(\sum_{i=1}^{i=n} \mathrm{P}(X_i = \mathrm{j}|\theta)\right).$$

Figure 4 displays this function on SSSB scale. TIMSS defined the cut score for the highest level as the point on scale where students are expected to Agree a lot (j=3) with four and Agree a Little (j=2) with three of the seven statements. $\mathrm{E(Total|\theta)}$ corresponding to this point 0*(0)+ 0*(1) + 3*(2) + 4* (3) = 18. The SSSB scale score corresponding to $\mathrm{E(Total|\theta)}$= 18 is 10.3 (Figure 4). TIMSS defined the cut score for the lowest level as the point on scale where students are expected to Disagree (j=1) with four and Agree a Little (j=2) with three of the seven statements, resulting in $\mathrm{E(Total|\theta)}$= 0*(0)+ 4*(1) + 3*(2) + 0* (3) = 10 and a SSSB scale score of 7.5 (Figure 4).

Table 2. Percentage of student endorsing each response option by trait level: Grade 8 TIMSS 2015 Students' Sense of School Belonging Scale items.

Trait level	Item	Disagree a lot	Disagree a little	Agree a little	Agree a lot	Total
Level3	I like being in school	1%	4%	50%	45%	100%
	I feel safe when I am at school	0%	2%	23%	75%	100%
	I feel like I belong at this school	0%	2%	23%	75%	100%
	I like to see my classmates at school	0%	1%	13%	86%	100%
	Teachers at my school are fair to me	0%	3%	27%	71%	100%
	I am proud to go to this school	0%	1%	19%	80%	100%
	I learn a lot in school	0%	1%	18%	81%	100%
Level 2	I like being in school	12%	26%	57%	6%	100%
	I feel safe when I am at school	5%	20%	55%	20%	100%
	I feel like I belong at this school	9%	26%	51%	14%	100%
	I like to see my classmates at school	3%	11%	41%	45%	100%
	Teachers at my school are fair to me	6%	23%	50%	20%	100%
	I am proud to go to this school	8%	25%	54%	13%	100%
	I learn a lot in school	2%	13%	56%	29%	100%
Level 1	I like being in school	66%	25%	9%	1%	100%
	I feel safe when I am at school	45%	34%	18%	3%	100%
	I feel like I belong at this school	77%	18%	5%	0%	100%
	I like to see my classmates at school	31%	24%	32%	13%	100%
	Teachers at my school are fair to me	51%	31%	15%	3%	100%
	I am proud to go to this school	74%	21%	4%	1%	100%
	I learn a lot in school	36%	35%	26%	4%	100%

Note. The highest percentage is printed with black color background.

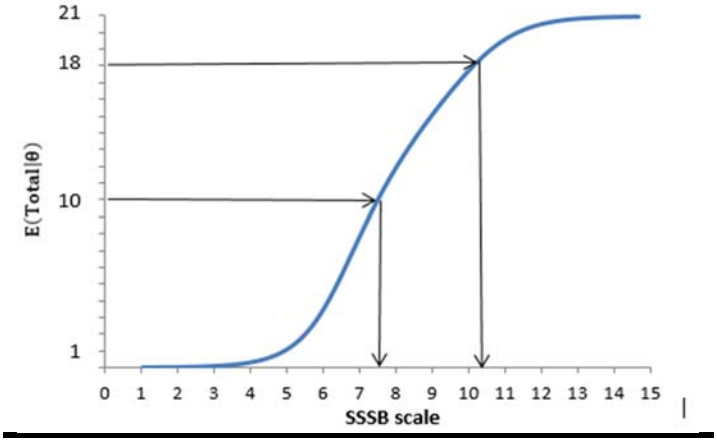

Figure 4. Expected total score curves for the Students' Sense of School Belonging Scale and associated cut scores identified by TIMSS: Grade 8 2015 assessment.

The difference between the TIMSS cut scores and those identified with the proposed approach are .52 (10.30-9.78) for the highest level (about 1/4 of a SD) and .64 (7.50-6.86) for the lowest level (about 1/3 of a SD). The percentage of students in the TIMSS US sample classified to the lowest, middle and highest levels according to cut scores identified by TIMSS and those derived using the approach proposed in this study were also computed. The Spearman correlation between the two classifications was .84 indicating relatively high correlation between classification results. The cut scores yielded relatively similar percentages for the middle level (49.0% vs 47.2%). TIMSS cut scores yielded higher percentage of students in the lowest level (14% vs 6.1%) and a lower percentage of students in the highest level (37% vs 46.6%). In addition, the correlation between trait level and TIMSS scores was .19 ($p<.05$) for both classification methods, for both mathematics and science.

Classification consistency and accuracy

Classifications based on any instrument measuring a latent trait are never error free: "Some examinees whose true ability is within a score range will have observed scores outside of that range." (Rudner, 2005, p. 1). This necessities the reporting of Classification Consistency (CC) and Classification Accuracy (CA). CC indicates the rate at which the respondent would be classified to the same category on two identical and independent administrations of the same measurement instrument. CA indicates the rate at which the respondents are classified to their true category. Therefore, CC relates to the reliability and CA relates to the validity of the classification (Lathrop, 2015).

CC and CA were examined with the approach Lee (2010) laid out. This approach is appropriate when the respondent ability and the cut scores are both on the total score metric. Since pattern scoring was not used in generating the SSSB scale scores, Lee's approach is appropriate in examining CC and CA[2] . Lee's CC and CA indices are based on the conditional observed score distribution derived with IRT models. The resulting distribution gives the probabilities of each total score for the examinee.

Based on the classification according to the cut scores identified with the proposed benchmarking approach, the overall CC (phi) and CA (phi) indices were 0.78 and 0.89, respectively. CC and CA for the classification based on TIMSS cut scores were 0.77 and 0.83, respectively. Therefore, the methods yielded similar CC and CA rates.

Discussion

In this study, an application of polytomous IRT models in identifying meaningful benchmarks on scales constructed with questionnaire items was illustrated. The major contribution of the approach is that it yields benchmarks with meaning/interpretation rooted in the item rating scale (Strongly Disagree to Strongly Agree). Given the item parameter estimates, the approach is relatively easy to implement. It can be applied to any scale based on items that are rated or endorsed on a common ordinal response scale as long as the items can be calibrated with a proper IRT model. Needless to say, such items must be conceptually related, all measuring the same underlying construct. As indicated earlier, model and item fit must also be investigated in advance.

The use of surveys such as the ones used in this study relate to group level results. The fact that standard errors for individual level scores might be relatively large when the scale of interest is based on a limited number of items does not diminish the value of such scales given that the main focus is almost always on the aggregated data. The use of group level results would be greatly enhanced if percentages of students at multiple meaningful levels can be reported and

[2] If pattern scoring is used in generating scores, the procedure discussed by Rudner (2005) would be more appropriate in examining CA.

compared as long as the CC and CA are high. The CC and CA of the resulting classification can be evaluated with procedure discussed by Rudner (2005) if scores are based on pattern scoring, or with the procedure discussed by Lee (2010) if scores are based on the total score metric. A major contribution of this study is the examination of CC and CA, which were improved compared to the more heuristic approach taken by TIMSS. Additional validity evidence, such as the one displayed in Table 2, should also be presented in future applications.

The approach illustrated in this study can also be used as a standard setting method for achievement tests that consist solely of polytomously scored items with identical number of score points for each. For example, in NAEP Writing assessments, a common holistic rubrics is used to rate student responses to all writing prompts on a seven point scale: 0 (unscorable), 1 (little or no skill), 2 (marginal), 3 (developing), 4 (adequate), 5 (competent) and 6 (effective). The application illustrated in this paper can be used to identify cut scores on this assessment where one or more of these seven score points are the most likely score. An investigation comparing the cut scores based on this approach for such an achievement test with those obtained using more traditional methods relying on subject matter expert judgment, such as the bookmark method, would also be a valuable contribution to the field.

Future studies could also investigate variations of the illustrated approach. For example, instead of identifying ranges on the scale where a particular response option is more likely than any other single response option, an alternative (or additional) criteria such as a set probability of endorsing the given response option (e.g. 75%) can also be sought. A comparative study between these variations would also be an informative contribution.

One limitation of the illustrated approach is that SCC for one or more of the response options might not be the most likely response for an acceptable range on the scale. This was the case for the *Disagree a little* option in this study. In such cases, defining the level in terms of likelihood of multiple response options might offer a solution as it was the case in this study.

References

Amtmann, D., Cook, K.F., Jensen, M.P., Chen, W.H., Choi, S., Revicki, D., Cella, D., Rothrock, N., Keefe, F., Callahan, L., & Lai, J.S. (2010). Development of a PROMIS item bank to measure pain interference. Pain, 150, 173–182. http://dx.doi.org/10.1016/j.pain

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Anthony, C. J., DiPerna, J. C., & Lei, P.W. (2016). Maximizing measurement efficiency of behavior rating scales using item response theory: An example with the Social Skills Improvement System–Teacher Rating Scale. *Journal of School Psychology*, 55, 57-69. doi:10.1016/j.jsp.2015.12.005

Bonanomi, A., Olivari, M. G., Mascheroni, E., Gatti, E., & Confalonieri, E. (2018). Using a multidimensional Rash analysis to evaluate the psychometric properties of the motivated strategies for learning questionnaire (MSLQ) among high school students. *Testing, Psychometrics, Methodology in Applied Psychology*, 25(1), 83-100.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed). Mahwah, NJ: Erlbaum

Carmichael, C. S., Callingham, R., Hay, I., & Watson, J. M. (2010). Measuring middle school students' interest in statistical literacy. *Mathematics Education Research Journal*, 22(3), 9–39.

Cook K.F., Victorson D,E., Cella D., Schalet B.D., & Miller D. (2015). Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. *Quality of Life Research*, 24(3), 575-89.

DiStefano, C. & Morgan, G. (2011). Examining Classification Criteria: A Comparison of Three Cut Score Methods. *Psychological Assessment*, 23 (2), 354–363.

Jong, C., Royal, K. D., Hodges, T. E., & Welder, R. M. (2015). Instruments to measure elementary preservice teachers' conceptions: an application of the Rasch rating scale model. *Educational Research Quarterly*, 39(1), 21–48.

Karatonis, A.,& Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4–12.

Lathrop, Q. (2015). Practical Issues in Estimating Classification Accuracy and Consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation*, 20(18). Retrieved September 10, 2017, from http://pareonline.net/getvn.asp?v=20&n=18

Lee, W. (2010). Classification consistency and accuracy for complex assessments using Item Response Theory. *Journal of Educational Measurement*, 47, 1-17.

Martin, M. O., Mulis, I. V., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (n.d.). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. Retrieved July 1, 2017, from https://timssandpirls.bc.edu/publications/timss/2015-methods/T15_MP_Chap15_Context_Q_Scales.pdf

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In: van de Linden W & Hambleton (Eds.), Handbook of modern item response theory. (pp. 101–122). New York: Springer.

Morgan, E. M., Mara, C. A., Huang, B., Barnett, K., Carle, A. C., Farrell, J. E., & Cook, K. F. (2017). Establishing clinical meaning and defining important differences for Patient-Reported Outcomes Measurement Information System (PROMIS(R)) measures in juvenile idiopathic arthritis using standard setting with patients, parents, and providers. *Quality of Life Research*, 26(3), 565–586.

Muraki, E. (1997). A generalized partial credit model. In: van der Linden W & Hambleton RK (eds.), Handbook of modern item response theory (pp. 153–164). New York: Springer.

National Center for Education Statistics (n.d.). Survey Questionnaires. Retrieved July 1, 2017, from https://nces.ed.gov/nationsreportcard/bgquest.aspx

Reeve, B. & Mâsse, L.(2004). Item Response Theory (IRT) Modeling for Questionnaire Evaluation. In Methods for Testing and Evaluating Survey Questionnaires , ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Rudner, L. M. (2005). Expected Classification Accuracy. *Practical Assessment Research & Evaluation*, 10(13). Retrieved September 1, 2017, from http://pareonline.net/getvn.asp?v=10&n=13

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement*, No.17.

Smith, E. V., Jr., Conrad, K. M., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement*, 10, 189-206.

Yovanoff, P., & Squires, J. (2006). Determining cutoff scores on a developmental screening measure: Use of receiver operating characteristics and item response theory. *Journal of Early Intervention*, 29, 48–62. doi:10.1177/105381510602900104.

Citation:

Dogan, Enis. (2018) An Application of the Partial Credit IRT Model in Identifying Benchmarks for Polytomous Rating Scale Instruments. *Practical Assessment, Research & Evaluation*, 23(7). Available online: http://pareonline.net/getvn.asp?v=23&n=7

Corresponding Author

Enis Dogan
Senior Education Research Scientist
National Center for Education Statistics at Institute of Education Sciences
U.S. Department of Education
550 12th Street SW, Room 4046
Washington DC 20202

email: enis.dogan [at] ed.gov