Some comments on representing construct levels in psychometric models

Ronli Diakow and David Torres Irribarra and Mark Wilson

Abstract This paper is concerned with one of the steps necessary to trace the connection between the substantive theory that serves as a basis for an assessment and the mathematical models that are used to analyze and rate student responses. We are interested in exploring this connection in the context of hypothesized variables that (a) have multiple ordered levels, (b) have been assessed with polytomous items that are meant to capture the aforementioned ordered performance levels, but (c) are modeled as continuous rather than ordinal. We present a straightforward method for estimated interpretable level boundaries when using the partial credit model. We then introduce graphical methods to evaluate the relationship between the levels as estimated by the model and the performance levels originally hypothesized by the theory. We believe that this kind of procedure can help practitioners make meaningful interpretations and provide more accurate diagnostic information to respondents in general.

1 Introduction

There is increasing interest in the development of *diagnostic* assessments that can provide actionable information to teachers to help them plan and target instructional activities. In order to successfully implement a system of assessments that provides such diagnostic information, it is necessary to coordinate cognitive theories about learning, our observations of student performance, and the interpretation

Ronli Diakow
University of California, Berkeley, CA, e-mail: rdiakow@berkeley.edu

David Torres Irribarra
University of California, Berkeley, CA, e-mail: dti@berkeley.edu

Mark Wilson
University of California, Berkeley, CA, e-mail: markw@berkeley.edu

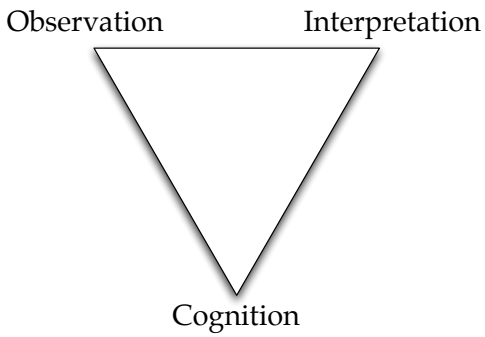


Fig. 1 The NRC assessment triangle.

of the evidence gathered during those observations. These are the components of the *assessment triangle* [17] presented in Figure 1. The alignment of these three areas is usually challenging, but is a necessary step to adequately embed meaning into the assessments used throughout the educational system.

This paper is concerned with one aspect of this process: tracing the connection between the substantive theory that serves as a basis for an assessment and the mathematical models that are used to analyze and rate student responses. We are interested in exploring this connection in the context of hypothesized variables that (a) have multiple ordered levels, (b) have been assessed with polytomous items that are meant to capture the aforementioned ordered performance levels, and (c) are modeled as continuous rather than ordinal.

1.1 Setting Performance Standards

Points (a) and (c) characterize the common case where the original theory that motivates the assessments classifies students into an ordered set of performances (e.g. "below basic", "basic", "proficient", "advanced") but the assessment models rely on the assumption of a continuous latent variable (e.g. Rasch Model, 2PL, 3PL). Traditionally when these conditions arise in the context of a criterion referenced assessment [9], stakeholders will use standard setting methods (see [4]) to provide cut-points in order to link the individual ability estimates back to the original levels of proficiency that the assessment was meant to differentiate.

It is worth noting that the term "standard setting" does not refer to a single procedure, but to a myriad of techniques [5], such as the *Bookmark Method* [13], the *Angoff Methods* [2] and its variations, and more recent *Holistic Methods* named in that way because "they require participants to focus judgment on a sample or collection of examinee work greater than a single item or task at a time." (p. 42) [5]. Overall the practice of standard setting has been described by Cizek as "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (p. 100) [3]. A specific method of standard setting has been developed

for the context where constructs have been designed based on identifying sequences of qualitatively different levels, such as the focus of this paper. The method is referred to as Construct Mapping [22], and is essentially a blend of the item-mapping concept behind the Bookmark method, and holistic methods. Recent examples are shown in work by Wilmot, Schoenfeld, Wilson, Champeney and Zahner [21] and Schwartz, Ayers, and Wilson [19].

1.2 Connecting Levels to Polytomous Tasks

The second point mentioned before, namely the use of polytomous items that are meant to capture the aforementioned ordered performance levels, is both more specific and less common than the other two. While many assessments are subject to a standard setting procedure, few of them have a strong connection between the theoretical levels of performance and the scoring procedures that would yield graded responses associated with specific performance levels.

We contend that the additional effort required for the development of items and scoring procedures with these characteristics can provide us with a good alternative to standard setting methods thanks to the alignment of the original theory, the assessment tasks, and the statistical model. In this paper we use an empirical dataset to illustrate how we can achieve traceability from the meaning based on substantive theory to the mathematical model used to estimate each student's location.

We address the overall question about how to connect the original theory (which contains levels) to the assessment model by answering three more specific questions:

1. How to delimit the boundaries between the levels (i.e. set cut-points)?
2. How to characterize the respondents whose estimates lie within a level?
3. How to evaluate if the estimated levels are consistent with the theoretical levels?

We answer the first question by illustrating a simple method for estimating interpretable level boundaries. We then introduce some graphical methods that can help practitioners address the second and third questions, namely, the interpretation of the performances associated with each level and the comparison to the levels originally predicted by the theory.

2 The Empirical Illustration

The empirical data for the illustration of the proposed methods comes from a larger multi-year project to assess adolescent literacy. The *Striving Readers* project focuses on a literacy intervention implemented by the San Diego Unified School District and funded by the Institute for Education Sciences called *Strategies for Literacy Independence across the Curriculum* (SLIC) [11, 16]. Students are taught how authors

use different text forms to present particular types of information and how the features convey information about the content of the text.

Construct Description		Scoring Guide
4	**Synthesizing - Creating New Key Ideas** - new understanding based upon the text - new understanding based upon multiple texts - evaluating author's intent - literary and/or rhetorical criticism	Response is **complete in relation to the information contained:** *Example*: This article is convincing you to get healthy by describing a program of exercise and eating right. It also encourages to do exercise and suggests that it is not hard.
3	**Cross-checking - Coordinating Key Ideas in the Text** - claim - argument - theme - identifying author's intent	Student responds with **multiple items from tactics-based sources and cross-checks or combines items of information:** *Example*: This article is about convincing us to stay fit and healthy.
2	**Discriminating - Key Ideas in the Text** - idea structure - supporting statement - plot - characterization	Student responds with **at least two items from tactics-based sources:** *Example*: Exercise is good for you. We should all exercise.
1	**Engaging - Ideas in the Text** - topic - main idea of a paragraph	Student responds with **one item of information from tactics-based source:** *Examples:* - A fitness plan - Exercise is good for you - Everyone should exercise - Changing your diet can enhance fitness results
0	**Disengaging - Ideas Not in the Text** - not challenging existing knowledge - no new ideas	Student gives an **incorrect response:** *Example*: It's about how you should exercise

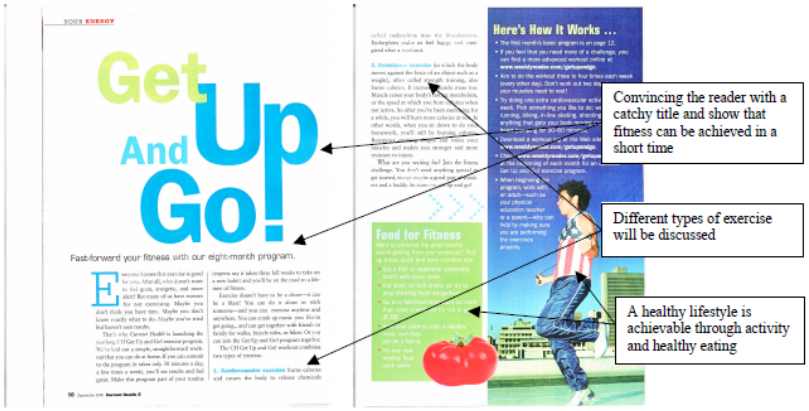
Fig. 2 The *Striving Readers* construct map and the scoring guide for one item.

A team from the Berkeley Evaluation and Assessment Research (BEAR) Center developed a system of assessments that would be embedded in the curriculum and also would be used as part of the evaluation of SLIC. The assessment development and refinement followed the Bear Assessment System (BAS) [23]. A full description

of the development of the assessments can be found in the project technical report [7].

Overall, 16 different assessments were developed for the SLIC curriculum, spanning the four grades covered by the curriculum (7th through 10th) and four assessment times in each school year (September, December, March, and June). Each assessment asks students to respond to a different text, and the genre of the text varies across the assessments. The same item types (questions that are similarly worded and address the same topics, e.g. main idea or inference) are used across the 16 texts as appropriate. Both the genres of text and the item types are a direct reflection of the instructional strategy implemented in the SLIC curriculum. All items on the SLIC assessments are scored polytomously. Since each of the 16 assessments relies on a different text, there are technically no common items across the assessments. A calibration sample, obtained in New Zealand[1] in the summer of 2008, was used to link the assessments; students in grades 7-10 took the assessments in an overlapping design that allowed linking through common persons.

The construct, assessments, items, and scoring guides were developed and refined jointly by the curriculum developers, district personnel, and assessment team. All 16 assessments were designed to assess the same underlying construct (Figure 2). The final construct had five levels, corresponding to increasingly sophisticated levels of reading comprehension. The literature suggests that the comprehension of written texts requires understanding the ways textual forms present particular types of information. The construct contains successive levels of identifying both how a text works and what the text means.



2. Scan the text features. What do you think this text will be about?

Fig. 3 A *Striving Readers* sample item.

[1] The calibration required that the assessments be taken by students who had used the new curriculum but were not potentially part of the experimental study in San Diego; schools in New Zealand, where the curriculum was first designed, were used.

Each assessment consisted of an authentic text (i.e. a published text, not written by the test developers) and 10-12 open-ended questions. Figure 3 shows an example copied from the scoring guide for one of the 7th grade assessments; it displays a portion of the text and one question. The text for this assessment is a persuasive text (a magazine article on the benefits of exercise) and the second question asks students to anticipate the content of the article based on the text features. The figure also points out the relevant text features for the rater (these hints were not given to the students).

Each response was scored from 0 to 4 with the score categories corresponding to one of the construct levels. Figure 2 shows another part of the scoring guide for the example item. Note how the score categories on the left side of Figure 2 match the construct levels on the left side of the same figure. The tests were scored by the teachers, curriculum developers, and other district personnel.

The subset of data used in this article comes from the calibration sample collected in New Zealand the summer of 2008. It consists of the responses of 202 7th graders with complete data for the 12 items of the 7th grade assessment using the persuasive text shown in the above example.

3 The Analysis and the Levels

The *Striving Readers* example nicely illustrates a construct that has a well defined set of ordered performance levels and a set of assessment tasks and scoring guides directly informed by those levels. However, it is in the final step, the measurement model used for analysis, that this connection is usually lost.

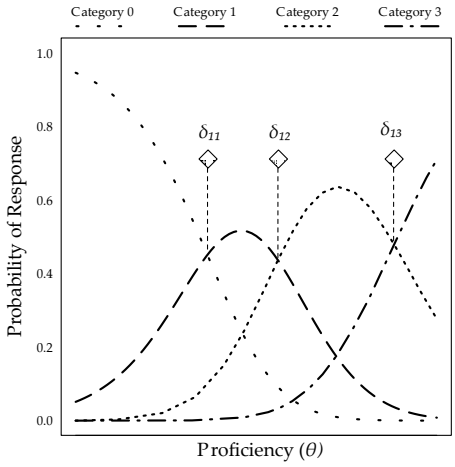


Fig. 4 Sample illustration of the locations of δ_{ij} parameters in a category characteristic curve plot.

A standard approach would be, for instance, to use a model like Master's Partial Credit Model (PCM) [15], where the logit of the probability of person p of answering item i at level j is:

$$\text{logit}[Pr(x_{pij} = 1 \mid \theta_p)] = \eta_{pij} = \theta_p - \delta_{ij} \tag{1}$$

In this model, the person proficiency is represented by θ_p, which indicates the person's location on the latent variable. Similarly, the difficulty associated with each category j in item i is represented by the δ_{ij} term.

This parameter represents the point on the latent variable when category j becomes more likely than category $j-1$. As we can see in Figure 4, δ_{11} is located at the intersection of the category characteristic curve (CCC) for category 0 and category 1, and similarly, δ_{12} and δ_{13} are located at the intersection of the CCC's of the respective adjacent categories.

As represented in the left panel of Figure 5, each δ_{ij} represents the interaction between the item and the category level, which means that the "levels" that are differentiated are effectively item specific and not common across the items. An alternative way to express the PCM, presented in the right panel on Figure 5, illustrates this more clearly by decomposing the interaction term δ_{ij} into the main effect of the item $\delta_{i.}$ (represented by the black diamonds) and an interaction term τ_{ij} (represented by the dashed lines).

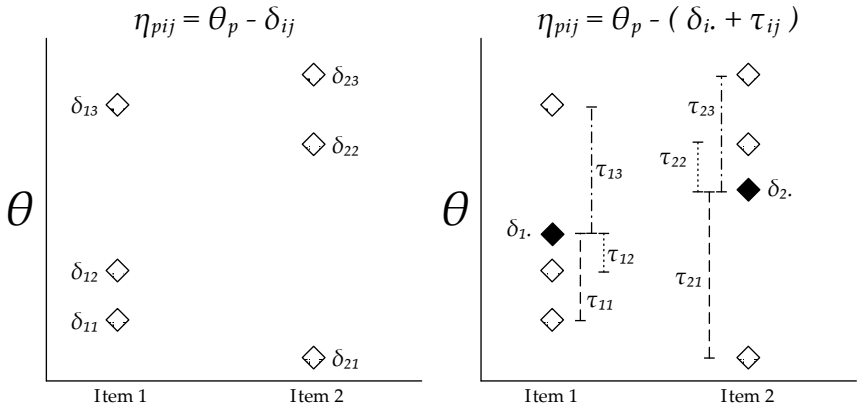


Fig. 5 Sample illustration comparing Master's parameterization of the PCM and a "rating scale" parameterization of the same model.

When the PCM is expressed in this way it is possible to see that there is no parameter associated with the level main effect, which effectively means that the original levels specified in the substantive theory are no longer directly represented as model parameters in this traditional parameterization of the PCM.

The PCM, like other polytomous logit models such as the graded response model [18] or continuation ratio models such as the sequential model [20], estimate these

interaction terms. They provide a better model fit than more restrictive models such as Andrich's Rating Scale Model (RSM) [1]. Instead of including an interaction term, the RSM relies on the use of main effects for both items and levels; however, the parsimony of the RSM often proves too restrictive, and the PCM is then preferred for the better fit it provides. For example, in the *Striving Readers* example, the PCM fit the data significantly better than the RSM ($\chi^2_{(22)} = 207.040, p < 0.001$).

It is possible to obtain the benefits of an overall level main effect (i.e. a $\delta_{.j}$ parameter) without resorting to the RSM by simply reparameterizing the PCM accordingly:

$$\text{logit}[Pr(x_{pij} = 1 \mid \theta_p)] = \eta_{pij} = \theta_p - (\delta_{.j} + \lambda_{ij}) \tag{2}$$

This simple reparameterization is graphically represented in Figure 6. It includes a overall level parameter $\delta_{.j}$ (represented by the dashed horizontal lines) that directly estimates a location for each boundary between the levels. The λ_{ij}, estimated as deviations from the level effects, are constrained such that the sum of the $\lambda_{i}j$ for a given level j is 0.

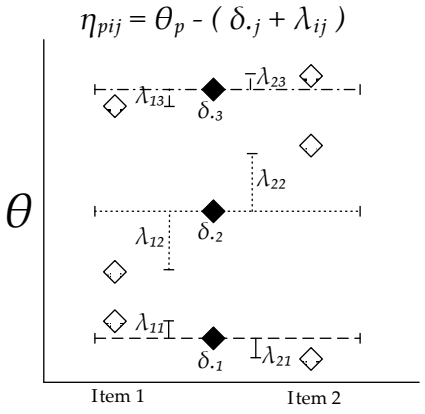


Fig. 6 Sample illustration illustrating the L-PCM reparametirization of the PCM including level main effects.

Using this alternative parameterization (henceforth L-PCM for level-PCM), we can preserve the link to the original theory and examine how closely the interaction terms λ_{ij} follow the overall level parameters $\delta_{.j}$ by examining their dispersion as shown in Figure 7. The item map in the figure organizes the λ_{ij} parameters around the overall level parameters, and the amount of observed variation in each level provides information about the quality of the overall levels as predictors of the item difficulty. In this example we can see that, although the dispersion of the λ_{ij} is considerable in all three levels, it is possible to appreciate a reasonable degree of separation between the three cluster of item by category parameters.

We have addressed the first of the questions raised on the introduction, namely how to estimate the boundaries (i.e. cut-points) between the levels. Based on these cut-points we can separate persons into the different levels of proficiency, fulfilling at least one of the goals that are usually pursued in standard setting exercises. How-

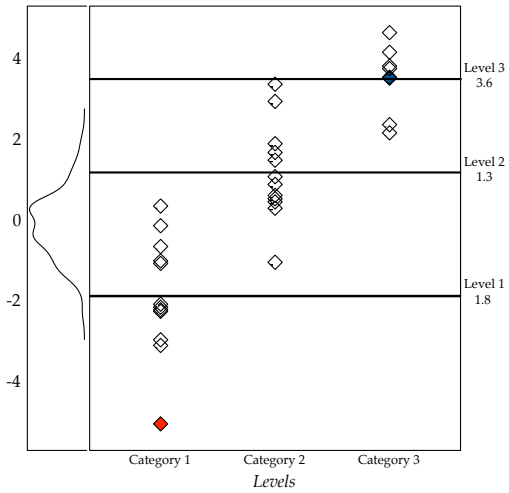


Fig. 7 Wright map organizing the λ_{ij} parameters as deviations of the $\delta_{.j}$ level parameters.

ever, it is one thing to determine cut-points, an another entirely different is to claim that those newly created levels in the latent variable correspond to the originally hypothesized performance levels.

4 Interpreting the Levels

Evaluating whether the estimated levels can be considered an accurate reflection of the theoretical levels is a critical issue that needs to be addressed in order to make tenable inferences about the respondents. We propose two complementary approaches to examine the "fit" of the estimated levels to the original hypothesis.

4.1 Identifying Ideal Cases

The first alternative for characterizing the proficiency levels estimated by the model is to consider them in relation to "ideal" cases, by which we mean response patterns that would canonically represent a performance level according to the original theory and task design.

We focus on two kinds of ideal cases, namely, prototypical and boundary cases:

1. **Prototypical cases:** refer to response patterns that would be considered exemplary of a given level, for instance, a respondent at level three that answers all the items at the level.

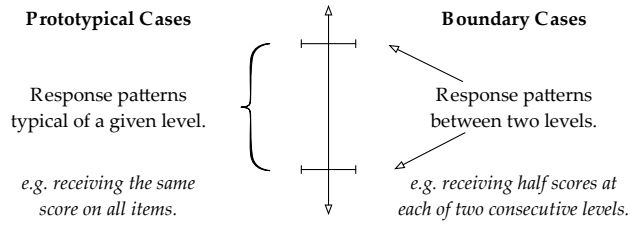


Fig. 8 Illustration of both kinds of ideal cases: prototypes and boundary cases.

2. **Boundary cases:** refer to response patterns that would be prototypical for a respondent that is "in between" two levels. For example, a respondent that is transitioning from level one to level two may answer exactly half of the tasks at level one and half of the tasks at level two.

Using these ideal cases as a guide, we can produce a plot, an example of which is shown in Figure 9, that "crosses" the average item scores of each one of these cases with the cut-points estimated directly by the model.

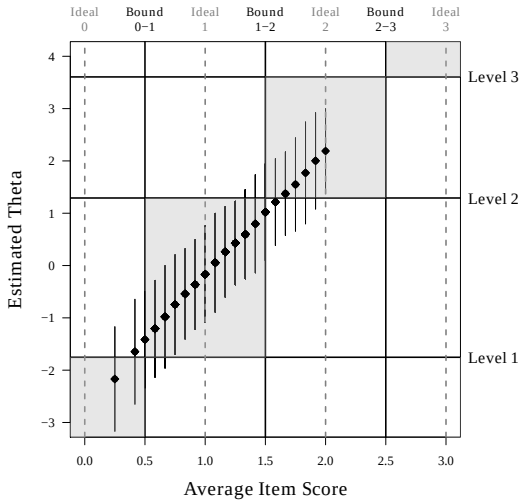


Fig. 9 Plotting proficiency estimates in relation to estimated cut-points and ideal cases.

For example, the prototypical case of a respondent at level three, answering all the items at level three, would have an average item score of 3 and the boundary case of a respondent transitioning from level one to level two, answering exactly half of the tasks at level one and half of the tasks at level two, would have an average item score of 1.5. In this plot, the "prototypical" cases (shown by gray dashed vertical lines) should indicate the centroid of the level, while the "boundary" cases (shown by black solid vertical lines) should signal the cut-points between the levels. The

horizontal black solid lines indicate the model estimates of the cut-points between the levels. Thus, the areas highlighted in gray indicate the regions where we expect our actual data to fall if the estimated levels match the hypothesized levels. We would compare the estimated person ability for each person[2] (given by solid black points) and their standard errors (the vertical grey lines around each black point) to the gray regions to check.

Figure 9 shows the results for the *Striving Readers* example. The unusual linear relationship between average item score and estimated theta is due to the even coverage of the sample ability range by the item parameters. Most of the points fall within the gray regions, indicating good fit between the hypothesized and estimated levels for this dataset. Perhaps the level two cut-point is slightly overestimated and the level one cutpoint slightly underestimated, since the data falls slightly outside the highlighted zones at those boundaries. Overall, the estimated level cut-points by the model appear to match the hypothesized levels.

This plot shows the zones in which we would expect the respondents to be located if there is a good match between the levels demarcated by the estimated cut-points and the levels hypothesized by the construct and allows us to compare the estimated person locations to those zones. Information about both kinds of ideal cases (Figure 8) is used to provide evidence regarding the match between the hypothesized and estimated levels.

4.2 Examining Expected Score Ranges

The second alternative for characterizing the proficiency levels that are being demarcated by the model cut-points is to examine the expected responses for the items within the range of each level. The expected item score for each item i is a function of θ_p and can be calculated by:

$$\sum_j jPr(x_{pij} = 1 \mid \theta_p) \tag{3}$$

using the estimated values for the item parameters.

Figure 10 shows the expected item scores for the 12 *Striving Readers* items from the 7th grade persuasive assessment. In order to examine the expected scores for a given level, we could use the area under each curve within that level, for example by integrating each curve in the left-hand side of the figure over each of the four estimated levels. Alternatively, we could examine the expected score for one (or more) discrete values of theta, as in the right-hand side of the figure. The two methods will yield similar results unless the expected score curves are irregularly shaped, and

[2] Note that when fitting a partial credit model on complete data, each average item score is a sufficient statistic for ability estimate; the figure would be messier but interpreted the same way if incomplete data were used.

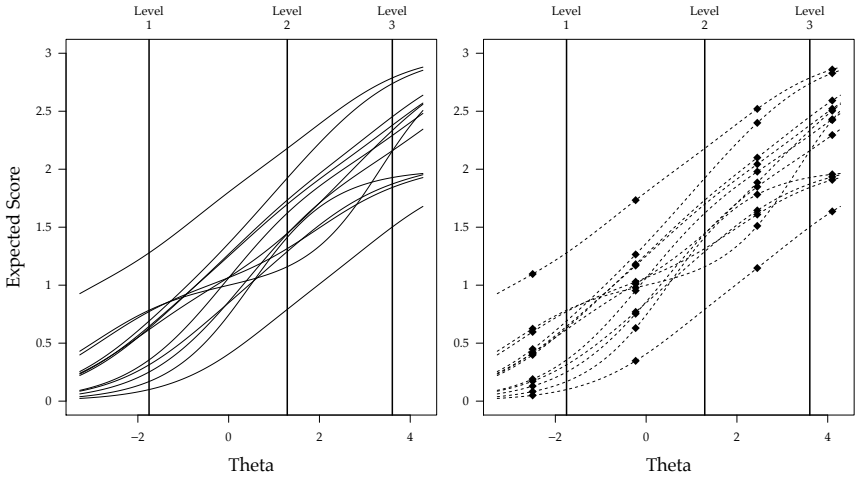


Fig. 10 Two approaches for calculating the expected item scores: area under the curves (left) or typical discrete values (right).

the latter method has the advantage of simplicity in calculation, presentation, and interpretation.

If we look at the range of the expected item scores associated with level two (i.e. between the 1-2 and the 2-3 cut-points), we would anticipate that the expected scores on that range are (a) aggregated closely around the value associated with a prototypical case and (b) within the range established by the boundary cases. An example of the kind of graph to do this is presented in Figure 11.

In this figure, the gray dashed vertical lines indicate the expected score for prototypical cases at each level and the solid black horizontal lines indicate the expected score for boundary cases between each successive pair of levels. The horizontal black solid lines indicate the model estimates of the cut-points between the levels. Thus, the areas highlighted in gray indicate the regions where we expect our actual data to fall if the estimated levels match the hypothesized levels. The black solid points give the expected item score for each item i at a given θ_p that is representative of the level[3].

Figure 11 shows the results for the *Striving Readers* example. In general, the expected item scores are centered around the gray dashed lines within each level and fall within the gray regions, indicating good fit between the hypothesized and estimated levels. There is possibly one item that is too easy and one that is too difficult; these items could be removed from further analysis. Note that there is not much data regarding level three, so in this particular example, we would conclude

[3] The representation of standard errors in this figure is not straightforward considering that the standard errors in the logit scale would be presented as horizontal bars, which would not contribute to the inference in the plot

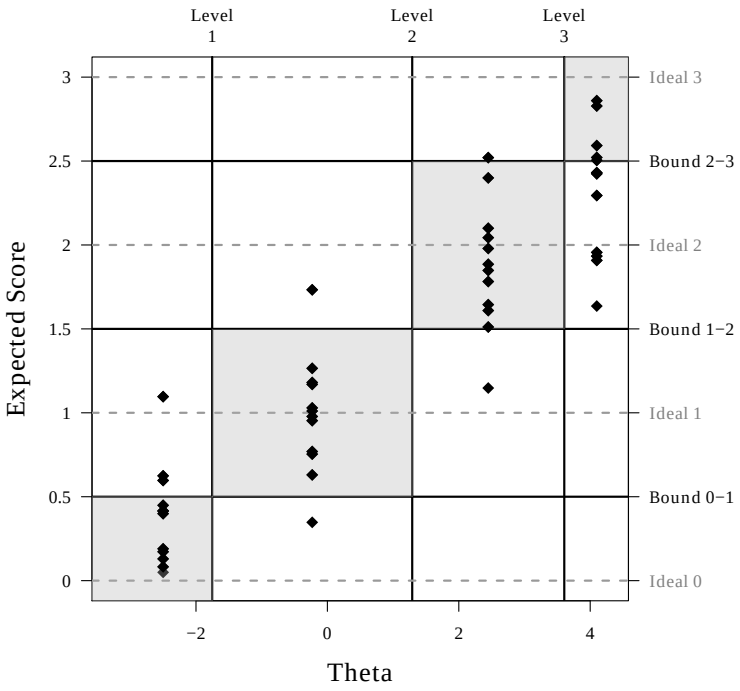


Fig. 11 Plotting proficiency estimates in relation to estimated expected scores.

that the hypothesized levels zero, one, and two are well recovered by the model but we would not make strong statements about the location of level three.

The two approaches, identifying ideal cases (as in Figure 9) and examining expected score ranges (as in Figure 11), provide complementary but distinct evidence. The first approach uses evidence from the person side of the model (the estimates of the person ability), while the second approach uses information from the item side of the model (the estimates of the item parameters). Since the person and item estimates are connected, the two figures will provide consistent evidence of the match between the hypothesized and estimated levels from alternative views of what the levels mean. However, we are investigating if one method is preferred in specific circumstances.

5 Discussion

5.1 Next Steps

Splitting a continuous variable into proficiency groups and then interpreting those groups, as is done in standard setting, is common practice. The methods to do so discussed in this paper could be extended by the use of located latent class models [14, 8], which directly model proficiency groups as in latent class analysis [12, 10], and a similar reparametrization to estimate a level main effect:

1. Located Latent Classes (LLC)

$$\eta_{pij} = \theta_{c(p)} - \delta_{ij}$$

2. Level Located Latent Classes (L-LLC)

$$\eta_{pij} = \theta_{c(p)} - (\delta_{\cdot j} + \lambda_{ij})$$

These models could be used to directly estimate the centroids of each level, represented by the $\theta_{c(p)}$, concurrently with the level cut-points. This would provide an alternative to the use of prototypical score patterns to characterize each level.

A second area where the direct estimation of level cut-points could be useful is in the context of previous work on Structured Construct Models (SCM) [24, 6]. In SCM, the relations between constructs are specified as conditional relations between specific levels of two different constructs. In other words, to reach level 4 of my "target" construct, the respondent is not only required to be at level 3 of that same construct, but also achieve a specific level, say level 3, of a "requirement" construct. So far, these models have relied exclusively on the use of latent class models to allocate the respondents into the different proficiency levels, but the estimation of the cut-points directly within the item response model opens the possibility of modeling these relations directly on a continuous latent variable.

5.2 Summary

In this paper we have presented a simple idea for improving the connection between the substantive theory used to create an assessment and the measurement model used to analyze it. Additionally, we have introduced two graphical representations that can help characterize the groups that have been established by the estimated cut-points by contrasting the observed performances of those groups to the performance levels originally hypothesized.

We believe that this kind of procedure can help practitioners make meaningful interpretations and provide more accurate diagnostic information to respondents in general. Furthermore, the procedure described in this paper can provide additional

evidence to support, corroborate, or potentially raise questions about the results of traditional standard setting procedures.

It is important to note, however, that the methodological simplicity of this procedure requires a considerable amount of work in the definition of the construct, the creation of the assessment tasks, and the elaboration of the scoring guides. It is our hope that the possibility of generating this kind of analysis will encourage test developers to invest effort in those earlier stages in order to improve the quality and interpretation of their assessments.

References

1. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.
2. Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.
3. Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement*, *30*(2), 93–106.
4. Cizek, G. J., & Sternberg, R. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.*. Mahwah, New Jersey: Lawrence Erlbaum.
5. Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting Performance Standards: contemporary Methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50.
6. Diakow, R., Torres Irribarra, D., & Wilson, M. (2011). *Analyzing the complex structure of a learning progression: structured construct models.* Paper presented at the annual meeting of the national council of measurement in education, New Orleans, LA, April 2011.
7. Dray, A. J., Brown, N. J. S., Lee, Y., Diakow, R., & Wilson, M. (2011). *Striving Readers BEAR Assessment Report.* Berkeley, CA: Berkeley Evaluation and Assessment Research Center.
8. Formann, A. K. (1995). Linear logistic latent class analysis and the Rasch model Rasch models: Foundations, recent developments, and applications. In G. H. Fischer & I. W. Molenaar (pp. 239-256). New York: Springer-Verlag.
9. Glass, G. V. (1978). Standards and Criteria. *Journal of Educational Measurement*, *15*(4), 237–261.
10. Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.*. New York: Cambridge University Press.
11. Institute for Education Sciences (2006). Striving Readers Program. Retrieved from http://www2.ed.gov/programs/strivingreaders/index.html. Cited December 10 2012.
12. Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* New York: Houghton, Mifflin.
13. Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: Abook- mark approach. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
14. Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96-107.
15. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.
16. McDonald, T., Thornley, C., Staley, R., & Moore, D. W. (2009). The San Diego Striving Readers' Project: Building Academic Success for Adolescent Readers. *Journal of Adolescent & Adult Literacy*, *52*(8), 720–722.

17. National Research Council (2001). *Knowing what students know : the science and design of educational assessment.* Washington, DC: National Academy Press.
18. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.
19. Schwartz, R., Ayers, E., & Wilson, M. (2011, July). *Mapping a learning progression using unidimensional and multidimensional item response models*. Paper presented at the International Meeting of the Psychometric Society, Hong Kong.
20. Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43(1), 39-55.
21. Wilmot, D.B., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning* , 13(4), 259–291.
22. Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000), pp 325–332. Tokyo: Springer-Verlag.
23. Wilson, M. (2005). *Constructing measures: An item response modeling approach.* New York: Routledge.
24. Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730.