The Basis of Scale Anchoring in Item Mapping: Some Issues of Concern

Andrew Kolstad
P20 Strategies LLC

Paper prepared for presentation to an Expert Panel convened by the Human Resources Research Organization (HumRRO) under contract number ED-NAGB-17-C-0002 with the National Assessment Governing Board, Arlington, Virginia, on July 12–13, 2018

The research reported here was partially supported by employment at the National Center for Education Statistics, U.S. Department of Education, during the period 1994-2004. The opinions expressed are those of the author and do not represent views of NCES or the U.S. Department of Education, nor those of the American Institutes for Research.

The author thanks his former colleagues Jon Cohen and Tsze Chan of the American Institutes for Research for their help with the early stages of this project and for providing essential data tabulations from restricted-use NAEP data files.

THE NAEP AUTHORIZING LEGISLATION, STILL IN EFFECT, continues to require that achievement levels be used on a trial basis until the Commissioner of Education Statistics determines that the achievement levels are “reasonable, reliable, valid, and informative to the public.” The legislation also requires that the Commissioner base this determination on a Congressionally mandated evaluation by a nationally recognized evaluation organization, such as the National Academies of Science, Engineering, and Medicine. Until that determination is made, the law requires the Commissioner and the Governing Board to state clearly the trial status of the achievement levels in all NAEP reports. The recent report of the Committee on the Evaluation of NAEP Achievement Levels for Mathematics and Reading (2017) recommended that the Commissioner remove this trial status when the achievement level descriptions have been updated to reflect current frameworks, assessment specifications, and items (while retaining the current cut scores).

In addition, the Governing Board is reconsidering its policy on developing student performance levels for NAEP (NAGB 1995). One aspect of its reconsideration is developing reporting achievement level descriptions. The current achievement level descriptions are developed as part of standard-setting activities with new or revised frameworks before and student performance data are available and are used initially to set achievement level standards. The new reporting ALDs would be developed using empirical data on student performance and would describe what students at each achievement level do know and can do, rather than what they should know and should be able to do.

Achievement level descriptions, in their preliminary and (currently) final forms, describe what students ought to be able to do to qualify for the three achievement levels (Reckase 2000). On occasion, ALDs have been derived from a scale anchoring process that ties the achievement level descriptions to what students do when responding to NAEP’s cognitive questions. By locating items requiring similar levels of knowledge together along a scale, subject matter experts can examine those that lie within selected scale ranges to identify the cognitive demands of the scale in that range, in a process known as ‘scale anchoring.’ The scale anchoring process was first used in the early days of NAEP, prior to the establishment of the National Assessment Governing Board (Beaton and Allen 1992, Mullis and Johnson, 1994, Johnson, *et al*. 1994). The Governing Board has also used scale anchoring processes to create achievement level descriptions for the 1996 science assessment (Bourque 1997), the 2005 and 2009 grade 12 mathematics assessments (with a new or revised framework), and the 2009 reading assessment (with a new framework and a large and complex bridging study). In each case, the scale anchoring process begins with item mapping (Lazer, *et al*. 2001).

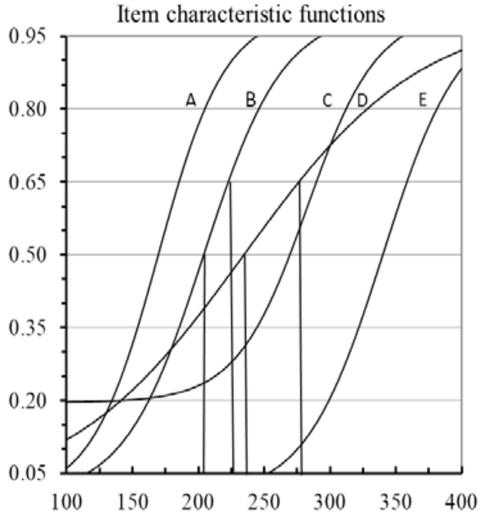
This paper has a narrow focus on a fundamental technical matter—those aspects of item response theory that bear on item mapping, a process that is critical both to bookmark methods of standard setting and to scale anchoring methods of developing achievement level descriptions. Item response theory offers a methodology for describing both examinee ability and assessment item difficulty in the same terms. Because an IRT model makes examinee ability and item difficulty commensurable, examining the content of items with similar difficulty has become the basis for creating descriptions of student performance in ranges along a cognitive scale. With item response theory, an examinee's test scores can be interpreted in terms of the types of test items that they can and cannot successfully perform.

Item mapping has at least two advantages in describing student performance. First, mapping items to scores shows how students would have performed on a limited set of publicly released items. This approach meets the simultaneous requirements for test security and accessible examples of student performance. Second, the mapping of exemplar items demonstrates, as examples of the types of knowledge and skills assessed by the cognitive scale, that policymakers, educators, researchers, and parents are interested in a domain of subject matter content, not just the exemplar items themselves.

Item mapping also has some limitations as a basis for developing descriptions of achievement within ranges along a cognitive scale. First, mapped items taken from any particular assessment year are but one of many possible samples from the universe of content defined by the framework. Over time, items consistent with the content framework are released to the public and dropped from future assessments, while different items are introduced to replace them in future assessments. Item performance is correlated with, but not identical to the cognitive trait. The item maps serve as one kind of input to subject matter experts as they try to balance the generalities of a content framework and the specific content of items that are sampled from the framework (Forsythe, 1993). Second, item difficulty can be idiosyncratic, attributable to the success of foils and distractors, and inconsistent with expert judgement of the cognitive demands of the item. Subject matter experts need to recognize the underlying cognitive demands of items and, to the extent that they are able, ignore the idiosyncratic elements.

The psychometric IRT model for mapping test items to test scores is probabilistic (except for Guttman's 1950 scalogram method), so each ability level is associated with some probability of a correct item response. Low abilities are associated with low probabilities of a correct response, and higher abilities are associated with higher probabilities of a correct response. No single number for an item can summarize what the model says about response probabilities as a function of the latent trait. Nevertheless, it is often useful to characterize an item's difficulty with a single number. The item thresholds, the *b*s in the model, best convey the difficulty of the items, and are expressed in the same units as the cognitive scale. However, other measures of the difficulty of items are possible.

As Bob Mislevy once told NCES (1999), "The threshold also turns out to be the location on the scale where item responses provide the *most* information for estimating examinees' thetas. The mathematics behind this fact leads to a curious paradox. At any given point along the theta scale, we have for items with their *b*s at that point the *least* information about how an examinee with that theta would respond. … Given theta, we literally could not say less about how we'd expect an examinee to fare." In reporting results, audiences find it confusing to claim that an examinee is proficient when he or she can perform a certain task only half the time. Many data users intuitively think that a 50 percent success rate corresponds to chance performance. Data analysts can counter this intuition by first selecting a desired probability of a correct response, and from that, find the scale score associated with that probability for any given item. Nontechnical audiences may find that mapping items to another location more intuitively acceptable.



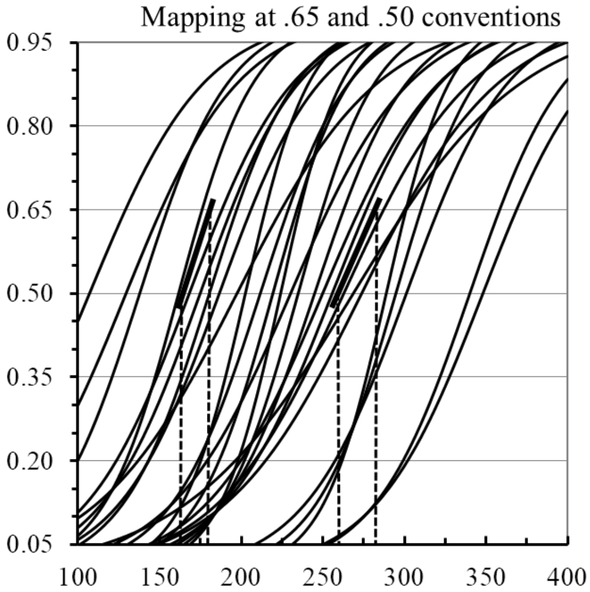
The concept of item mapping is illustrated in the following figure, where an IRT-based proficiency scale is shown on the horizontal axis on a NAEP-like scale ranging from 100 to 400 and the probability of a correct response is shown on the vertical axis. In the figure, five hypothetical items (labelled A to E) are plotted in the form of a two- or three-parameter logistic model. Over this range, the examinees show a gradually increasing probability of success. This graph adds horizontal guidelines at several locations along the probability axis—particularly important are those for which the probability reaches .50, .65, and .80. With cognitive items B and D, proficiency levels of 226 and 279 (respectively) are required to be able to succeed with a 0.65 response probability. With a 0.50 probability given by the item's threshold, proficiency levels of 205 and 235 are required. Because item D has a lower discrimination parameter than item B, the required proficiency levels are closer together for the 50 percent convention (30 points) than for the 65 percent convention (53 points).

The low predictability of success at the threshold led some analysts to associate a different point on the scale with an item, using a criterion called the *response probability convention.* High response probability conventions have typically been justified in terms of "mastery" (Bock, Mislevy, and Woodson, 1982; Beaton, 1987). The mastery argument is fairly simple: if one is going to say that people with a particular score on an assessment can successfully perform a particular assessment task, one wants to be fairly sure that a substantial majority of them can do it. Intuitively, this perspective is appealing because the notion of mastery carries with it an expectation that successful performance will be consistent. Bock, Mislevy, and Woodson (1982) describe a preference for high values for a response probability convention as follows:

> While the traditional practice in mental test theory is to define an item's threshold as the point at which pupils have a 50 percent chance of responding correctly, we believe it is preferable in assessment to discuss item content with respect to a higher degree of mastery indicating the level of skill at which a majority of students are functioning. We refer to the 80 percent point of an item as its 'mastery threshold.'

This intuition about item mastery implies that the claim that an examinee has the proficiency needed to succeed with a given test question should be supported by mapping the question onto the proficiency scale at a point that ensures that a substantial majority of examinees at that point on the scale will answer the question correctly.

Item mapping is widely used in bookmark-based standard setting methods. With this approach, items are arranged in ordered item booklets, using the scale score point associated with the chosen response probability convention. The task of judges is to place a mark between items in an ordered item booklet that distinguishes between the items that an examinee in a reporting category should be able to answer correctly and those they would not be able to answer correctly. Each judge would have a score associated with the point bordered by the two kinds of tasks, and the average of the judges cut scores becomes the basis for recommending cut scores to the responsible policy body.



The locations of those cut scores depend on the value of the response probability that has been adopted as a convention. In the graphic to the right, the judge has placed a lower bookmark between items 4 and 5 (counting from the left), which results in a cutscore of 175 at the 0.65 response probability and a cutscore of 162 at the 0.50 threshold score. The judge has place an upper bookmark between items items 7 and 8 (counting from the right), which results in a cutscore of 286 using the 0.65 response probability convention and a cutscore of 252 at the 0.50 item threshold score. The selection of a response probability convention affects the location of the standard-setting panelists' cut scores in a way that is rarely obvious to standard-setting policy boards (Kolstad, 1999). In the present context, for which the Governing Board intends to keep the reading and mathematics achivement level cut scores fixed, the effects of the response probability convention would appear not in the standard-setting process, but in the scale-anchoring process, for which item maps are likely to serve as input to subject matter expert panelists.

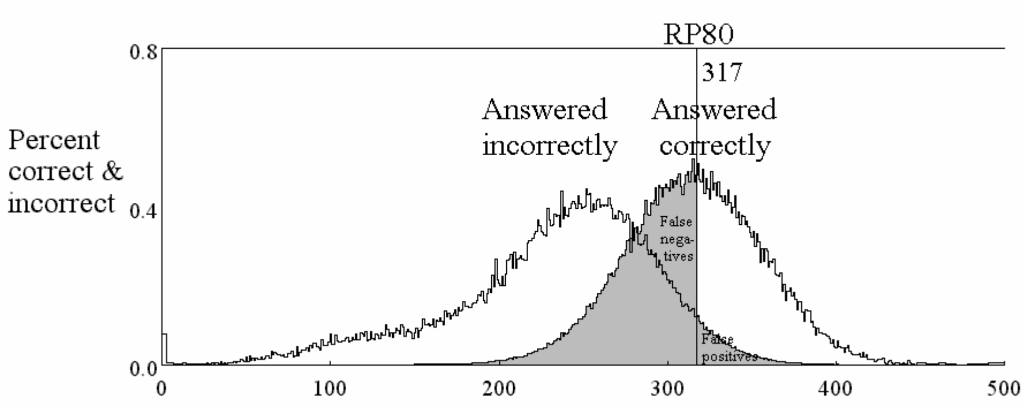
In my two decades of research and statistical activities with NAEP and the adult literacy assessments, I have studied item mapping and have identified several problematic aspects of the common intuition about item mastery. In this essay, I will describe my problems with the response probability convention and present several alternative approaches that address these problems. I hope that some of what I have found can be useful to the Governing Board as it considers how to revise the existing reading and mathematics achievement level descriptions to make them more consistent with student performance and how to develop reporting achievement level descriptions in the future that would be empirically based and useful for reporting.

Prescribing sufficient mastery. The basic intuition that brought about the use of a response probability convention with item response theory cognitive scales was the common-sense notion that 50/50 odds of success on an item are insufficient to indicate mastery of the content. Soon after ETS began to conduct NAEP and introduced IRT scaling to NAEP's first reading assessment scale—fielded in 1983-84—Al Beaton (1987) developed anchor level descriptions based on the admittedly arbitrary response probability convention of 0.80 (with a companion standard for sufficient item discrimination). His choice was influenced by the recent publication of the Bock, Mislevy, and Woodson (1982) paper that referred to 0.80 as the 'mastery threshold.' This convention carried over (without the item discrimination requirement) to the 1992 National Adult Literacy Survey, which used a block of grade 12 NAEP reading items and much of NAEP's assessment survey methodology (Kirsch & Jungblut 1986, Kolstad 2001). However, when the 1986 NAEP mathematics assessment tried the approach used for 1984 NAEP reading, too few items mapped at the top anchor level. In response, the ETS staff decided to split the difference between 0.80 and 0.50, and mapped the mathematics items at 0.65 (Johnson 1988, 1994). Again, when the National Assessment Governing Board tried to use the 0.65 convention for mapping exemplar items, they found few items that mapped at the Advanced level and reduced the criterion back to the IRT threshold of 0.50. Small differences in the response probability convention can make for large differences in items that map to a given level.

My first problem with this common conception is that simply claiming that a response probability of 0.50 is *not* sufficient provides no guidance for what *would be* sufficient. Hyunh (1998) decomposed the item information into that provided by a correct response and that provided by an incorrect response. Hyunh showed that the item information provided by a correct response to a constructed-response item is maximized at the point along the scale at which two-thirds of the students get the item correct. This sounds like a technical argument in support of the 0.65 response probability convention. However, maximizing all of the item information—taking into account both responses that are correct and those that are incorrect—would imply a response probability convention closer to 0.50 (every NAEP technical report describing the process of item mapping makes this point). While national assessment surveys have adopted various conventions over time in response to practical needs (e.g., Zwick *et al.* 2001), and while state assessment systems have often settled on 2/3 odds (a 0.67 convention) for ease of explanation to standard-setting panels, the fact remains that there is no definitive and theoretical guidance for how high the response probability convention should be.

Imbalance of false positives and false negatives. Using the mastery concept to map items succeeds in reducing the proportion of people who seemingly have enough skill, but answer the question incorrectly (false positives). However, there is a cost to be paid in that this mapping can't be done without increasing proportion of people who seemingly have insufficient skill, yet still answer the question correctly (false negatives).

The following figure illustrates this point with a real assessment item taken from the 1992 National Adult Literacy Survey prose literacy scale (Kolstad, 1999). Nearly half the adults in this

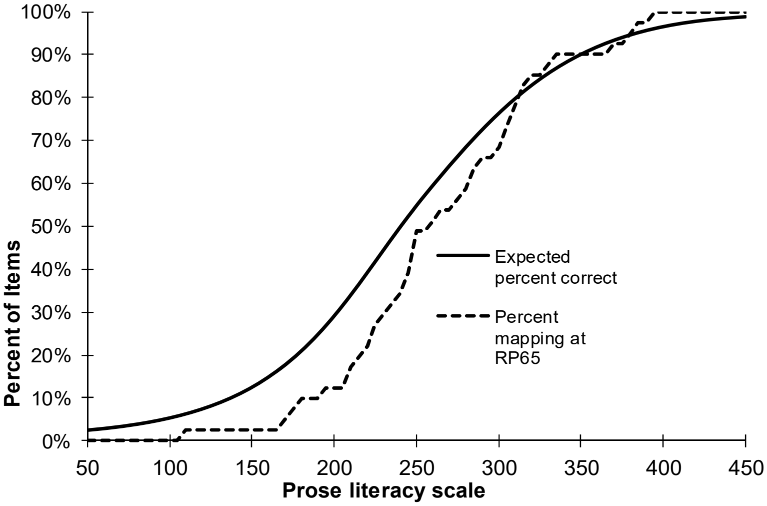


survey answered this item correctly. This item maps at 317 using the 80 percent response probability criterion. This figure shows two population distributions — the scale scores both for those who succeeded and for those who failed to answer this question correctly. Those who scored above 317 appear to have mastered the item, since only a very small the part of the distribution of scale scores for those with incorrect responses lies above 317. On the other hand, there is a good deal less confidence about the lack of ability on the part of those who score below 317. Nearly half of those who answered this item correctly apparently have too little prose literacy to be able to do so. While about half of the population answered this item correctly, only a quarter of the population scored above 317, indicating that they had sufficient skill to meet this particular mastery criterion. The item mapping convention demands that the audience dismiss the substantial proportion of adults with scores below 317 who answered correctly and interpret such success as accidental.

Stiffer standards for weaker questions. The early ETS use of scale anchoring required that anchor items (which amounted to about half of the total reading item pool) discriminated well between levels of performance. Using a response probability convention well above 0.50 results in a smaller difference between test questions that discriminate well and a larger difference between items that discriminate more poorly, as the first example above illustrated. Items that discriminate less well are less correlated with the latent cognitive trait, and should count less in describing performance along the scale.

An equivalent approach to achieving a high response probability would be unacceptable. In item response theory, the difference between the cognitive score θ, and the item threshold, b determines the probability of a correct response. Item mapping essentially adds a constant x to the item threshold, $\theta - (b + x)$, making the item appear more difficult. Mathematically, it is equivalent to subtracting the same constant from the cognitive score: $\theta - (b + x) = (\theta - x) - b$. As I estimated the size of this x, I found that score decrement to be applied to an entire examinee population in order to raise the response probability to any desired level (Kolstad & Wiley 2001). With the document literacy scale used in the 1992 National Adult Literacy Survey, raising the response probability from 0.50 to 0.65 would require subtracting 17 points from the scores of every adult (when the less discriminating items are excluded) and 20 points when all items are included. (The width of the reporting levels was 50 points.) To my knowledge, no testing system has used this method of attaining higher response probabilities, probably because reducing examinee scores would be unacceptable to any audience. While the two methods are mathematically equivalent, public audiences have not raised concerns about raising the score locations of cognitive items. Raising the score locations makes it appear that the population of examinees knows and can do less than item response theory thresholds for cognitive items would imply.

Disagreement with p-values: The proportion of the population that correctly answers each item is usually larger than the proportion that reaches any scale point with a response probability convention well above 0.50. This is not just a problem for a single item or for the 0.80 probability convention. Confusion is likely between the expected percentage of items answered correctly and the percentage who meet the 0.65 probability convention. The figure of the right shows the cumulative expected percent correct and the cumulative percentage of items that map at 0.65 for the 1992 National Adult Literacy Survey prose literacy scale. Except at the high end of the scale, the latter is generally too small as an estimate of the former.

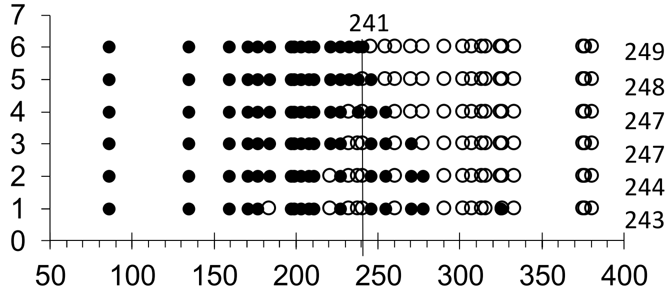


It is easy for the public to confuse achieving a scale score sufficient to ensure at least a 0.80 probability of a correct response on some particular item with simply getting a correct answer (Linn and Dunbar, 1992). For example, the *New York Times* reported that “a Federal study showed that almost half of American adults possessed very limited proficiency in English. In each category, about half of those tested could not answer a question more difficult than the one shown” (Celis 1993). The story then displayed three illustrative level 2 literacy tasks, one for each scale. The published report indicated the percentage of adults that scored high enough to reach the 0.80 probability of success on these items, but not the percentage of adults who correctly answered the three illustrative questions. The missing fact is that while only 52 percent of adults reached 275 or higher on the prose literacy scale, 72 percent of adults correctly answered the illustrative ‘appliance repair’ prose literacy task (and similar differences exist for the other two questions). The *New York Times* reporter clearly failed to understand the distinction between not being able to answer the question correctly and not having a scale score sufficient to ensure at least the 0.80 probability of success. The use of the 0.80 convention distorts the public perception of who is literate enough to correctly perform the tasks.

It is also possible to confuse the percentage of tasks answered correctly within a literacy level with the response probability criterion. Consider, for example, performance in level 3 on the National Adult Literacy Survey’s prose literacy scale (a score between 276 and 325) and hypothesize that some individuals responded to all the literacy tasks on the prose literacy scale. But someone who answered correctly all of the literacy tasks in levels 1 and 2, none of the literacy tasks in levels 4 and 5, and the easiest 80 percent (nine of eleven) of the level 3 tasks, would score only 268 on the prose literacy scale, too low to qualify for level 3. Thus, simply answering correctly 80 percent of the items within a level is not sufficient to be classified as performing in that level. Someone who answered correctly all of the tasks in levels 1, 2 and 3 but none of the tasks in levels 4 and 5 would score 277—just barely above 275, the lower boundary of level 3. To me this is not analogous to answering 80 percent of the questions on a test, but is analogous instead to answering more than 100 percent (because the examinee has to answer correctly more than the items in level 3 in order to be counted as performing in level 3).

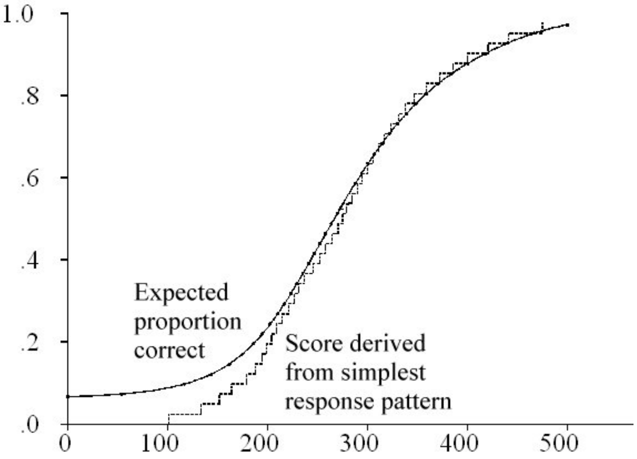
Considering items apart from scale scores: At an item's IRT-defined threshold value, *b*, one's intuition that the item has not been mastered can be misleading, unless one also understands the implications of a constant scale score. If the examinee actually answered the given question incorrectly, to have the same score, the examinee would have had to answer some other question correctly (or else the examinee's score would not have been constant). When items are mapped to a proficiency scale, it is intuitively sensible, particularly for the lay public, to expect that an examinee with a given score will correctly answer items mapped below that score, and incorrectly answer items mapped above that score. The resulting pattern of easier right/harder wrong response patterns can be described as a simple response pattern, and would be the response pattern expected if items on the scale followed Guttman's (1950) scalogram concept (Embretson and Reise 2000).

Consider the question of whether that other cognitive item would have to be harder or easier. The question of what happens when an examinee with an item mapped at 0.50 and a given score gets that item wrong can be illustrated by taking a random score pattern that produces a given score and changing the hypothetical answers to become a scalogram pattern. The figure to the right is a hypothetical assessment based on 29 selected items (with high discriminations and spaced as far apart as practical) taken from the prose literacy scale of the 1992 National Assessment of Adult Literacy. The bottom row shows a randomly generated response pattern based on a true score of 243, which is close to the threshold *b* value (241) of item N080201—an item of moderate difficulty. Correct answers are shown as dots, and incorrect answers are shown as circles. Next to each plot of a response pattern is its associated maximum likelihood score. Each successive row above switches answers to one pairs of items, as they approach the pattern of a scalogram at the top. In each pattern except the top one, the answer to item N080201 is incorrect. At the pattern just below the top, the answer pattern shows that target item was answered incorrectly, but one of the more difficult items was answered



correctly. This example illustrates that the only way a respondent can have the given score while incorrectly answering the item mapped at 0.50 is to answer a *more* difficult item correctly. This would not indicate a lack of mastery of the marginal item, but the substitution of an item with equivalent or greater difficulty. Ignoring the constancy of the score gets one's intuition into trouble.

The IRT maximum likelihood score calculated from the item parameters associated with simple response patterns can be associated with the most difficult correct response in the pattern. While items have no unique ordering, for this purpose local ordering can be used, which places items in the order that they would have near the point where a right or wrong answer in the response pattern matters most. The response probability associated with such a score can also be calculated. The figure to the right is based on the 41-item prose literacy scale from the 1992 National Adult Literacy Survey. With such a large number of items, the match between the expected number correct and the corresponding simple score pattern is fairly good, except in the lowest range of the scale. When I and my co-authors calculated the scores associated with such scalogram patterns for three NCES assessment surveys (1994 NAEP reading, 1992 National Adult Literacy Survey prose literacy, and the reading scale from the National Education Longitudinal Study of 1988, the average response probability for the marginal items across all patterns for each data set were very near 0.50, with the average for the 1992 National Adult Literacy Survey falling at 0.51 (Kolstad, *et al.*, 1998). [This occurred in part because that survey, unlike the other two, contained almost no multiple choice items which would tend to raise the average.]



In assessment surveys, there is little policy interest in whether examinees master particular items themselves. Rather the goal is to measure whether examinees master the underlying proficiency that the items are supposed to capture. Item response theory models have traditionally identified a response probability of 0.50 as the point that characterizes the difficulty of an item. This is due to IRT measurement drawing a clear distinction between all the skills required to successfully perform, for example, mathematics assessment task and the mathematical ability required to perform that mathematics task. Successful performance on such a task requires a specified level of mathematical proficiency plus other nuisance proficiencies, the latter being considered measurement error. Thus, when an individual's math ability exactly matches the mathematics requirements of an item, that individual has exactly a 0.50 probability of getting the item right. If one trusts the assumptions of item response theory modeling, a response probability of 0.50 should be used, since it maximizes the information obtained from the item.

In my opinion, there are four options for mapping cognitive items onto an assessment scale for input to subject matter experts who can write achievement level descriptions that reconcile performance on a sample of cognitive items with performance on the range of similar items that could be sampled from the same framework.

Option 1: Use a judgmental process to choose an increment to the IRT b parameter that takes into account the principal policy uses of the data. For some policy purposes, the balance between false positives and false negatives may differ from those for other purposes. For example, if special services such as literacy remediation are going to be targeted to low performers, we ought to be very sure that they need the services by setting the response probability convention below 0.50 rather than targeting setting it above 0.50 to ensure that examinees are more than capable of the tasks they are set. Since this is a policy decision that depends on the purpose for which the data are expected to be used, there is no reason to rely on the conventional practice of a 0.65 probability. However, it is difficult for assessment programs with multiple uses to focus its procedures on any one expected use. This approach also uses different criteria to place cognitive items and examinees on an assessment scale and produces a degraded correspondence between the percentage of questions answered correctly and the percentage of items that meet the response probability convention.

Option 2: Map items by matching the distribution of scores in the population and the p-value of the item. By assigning to the test question the scale score corresponding to the point on the latent distribution at which the percentage of the population achieving at least that point matches the percentage of the population that answers the question correctly. This approach matches the population distributions of success on cognitive items and success at points along the proficiency scale. However, this method uses different criteria to place cognitive items and examinees on the assessment scale, places more stringent standards on easy questions and less stringent standards on harder questions (compressing the items together along the scale). When dealing with nontechnical, content-expert panelists, this method might need an explanation to understand how the difficulty of the individual items corresponds to scale scores.

Option 3: Map items using simple response patterns. Under this option, each cognitive question would be assigned the scale score that would be received by an examinee if that question were the most difficult item answered correctly in a simple scalogram pattern of responses. This approach produces a good match between the mapping of items and the percentage of correct answers needed to qualify for the corresponding score. However, this method places less stringent standards on easy questions and more stringent standards on harder questions (spreading the items out along the scale). It uses similar, but not identical criteria to place cognitive items and examinees on the assessment scale. In my view, this method has an intuitive explanation that helps nontechnical panelists to understand why the probability of a correct response for an item on the margin is close to 0.50, yet the examinee possesses the ability to answer that or a similarly difficult question correctly.

Option 4: Map items at the IRT threshold parameter (without adjustment for guessing). This approach uses the same criteria to place cognitive items and examinees on the scale, places equally stringent standards on all items, produces a passable correspondence between the percentage of questions answered correctly and the percentage of items that meet the response probability convention, but when dealing with nontechnical, content-expert panelists, will need an explanation to counter the basic intuition about the lack of predictability about success with the individual items that they are responsible for examining closely.

Conclusion. Since the 1980s, over many NAEP and state standard-setting and scale anchoring activities, the 0.65 response probability convention has continued to be used. In my opinion, the attraction of this convention has been that the mastery concept remains intuitively plausible and persuasive to nontechnical, content-expert panelists. Until an alternative intuitive explanation can be found, the technical, IRT-based problems with this approach will have little influence over policy boards and over those who conduct standard-setting and scale anchoring processes.

My candidate for an alternative explanation that a 0.50 response probability, one that should be acceptable to nontechnical panelists, relies on the analogy to the simple response patterns of a Guttman scalogram. Any given pattern of responses has a set of answers that will rarely look like a scalogram, but can be turned into a scalogram by switching the correct answers to more difficult items with incorrect answers to easier items until a scalogram pattern is reached—one in which all easy questions were answered correctly up to the point determined by the scale score and all more difficult questions were answered incorrectly. The most difficult answer in a scalogram pattern of answers will have a response probability close to 0.50 (or somewhat higher if many multiple choice items are included in the assessment). Anyone who answered that marginal question incorrectly at a given scale score must have substituted a more difficult question, and thus can be considered to have mastered the content of the scale at that point.

The issue of item mapping may appear to be of interest only to technical staff, but because of its large impact on standard setting and developing descriptions of performance along ranges of a cognitive scale, it should be important to everyone. When the mastery intuition is accepted and items are mapped above their IRT-determined threshold, the performance of entire populations appears weakened and no anchoring or exemplar items can be found at the upper regions of the scale. The descriptions written by content experts of subsets of ability anchored to various ranges along a cognitive scale become underestimates of what examinees scoring in such ranges can do. The mastery concept builds bias into the underpinnings of scale score descriptions and of student performance.

References

Beaton, A. 1987. "The NAEP Reading Scale," Chapter 10.5 in *Implementing the New Design: The NAEP 1983–84 Technical Report*. Princeton, NJ: Educational Testing Service.

Beaton, A. E., and N. L. Allen. 1992 . "Interpreting scales through scale anchoring." *Journal of Educational Statistics*, 17: 191–204.

Bock, R. D., R. Mislevy, and C.Woodson. 1982 . "The next stage in educational assessment." *Educational Researcher*, 2 (March), 4–11, 16.

Bourque, M.L. 1994. "The NAEP achievement level-setting process for the 1992 mathematics assessment." Appendix G, pp. 841–861 in E.G. Johnson, J.E. Carlson, et al., *The NAEP 1992 Technical Report*. NCES 94–490. Washington, DC: National Center for Education Statistics.

Bourque, M.L. 1994. "The NAEP achievement level-setting process for the 1992 reading assessment." Appendix H, pp. 865–890 in E.G. Johnson, J.E. Carlson, et al., *The NAEP 1992 Technical Report*. NCES 94–490. Washington, DC: National Center for Education Statistics.

Bourque, M.L. 1997. "Report on developing achievement levels descriptions for the 1996 NAEP science assessment" Washington, DC: National Assessment Governing Board. Available on request.

Celis, W. 1993 . "Study says half of adults in U.S. can't read or handle arithmetic." *New York Times*. (September 9).

Committee on the Evaluation of NAEP Achievement Levels for Mathematics and Reading. 2017. *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress*. C.Edley, Jr., and J.A. Koenig, (eds.). National Academies of Science, Engineering, and Medicine. Washington, DC: The National Academies Press. doi: 10.17226/23409

Embretson, S.E., and S.P. Reise. 2000. "Measuring persons: Scoring examinees with IRT models." Chapter 7, pp. 158-186 in Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Forsyth, R. 1998. "NAEP frameworks and achievement levels." Pp. 3-26 in M.L. Bourque, (ed), *Proceedings of Achievement Levels Workshop*. Washington, DC: National Assessment Governing Board.

Guttman, L. 1950. "The basis for scalogram analysis." Chapter 3 in S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen, *Studies in Social Psychology in World War II: Volume IV, Measurement and Prediction*, Princeton: Princeton University Press.

Huynh, H. 1998. "On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation," *Journal of Educational Statistics and Behavioral Statistics*. 23: 35-36

Johnson, E.G. 1988. "Anchoring the points on the mathematics composite," Chapter 10.1.6, pp. 235–236 in Albert Beaton (ed.) *Expanding the New Design: The NAEP 1985–86 Technical Report*. Princeton, NJ: Educational Testing Service.

Johnson, E.G. 1994, "Description of percentages for anchoring and item mapping." Internal ETS memo, February 4.

Johnson, E.G., I.V.S. Mullis, J.R. Campbell, and S.P. Isham. 1994. "The NAEP scale anchoring for the 1992 reading assessment." Appendix J, pages 909–926 in Eugene G. Johnson and James E. Carlson (eds.), *The NAEP 1992 Technical Report*. NCES 94–490. Washington, DC: National Center for Education Statistics.

Kirsch, I. S., A. Jungeblut, *et al*. 1986 . "Describing and anchoring the scales" and "Levels of Proficiency." In *Final Report: Literacy: Profiles of America's Young Adults*, Pp. III-9/III-10 and IV-11/IV-13, Princeton, NJ: Educational Testing Service.

Kolstad, A. 1996. "The response probability convention embedded in reporting prose literacy levels from the 1992 National Adult Literacy Survey." Paper presented at the annual meeting of the American Educational Research Association, April 1996.

Kolstad, A. 1999. "Standard-setting by the back door: 'Mastery' as a criterion for mapping items onto IRT scales." Paper presented to the CCSSO National Conference on Large-Scale Assessment, Snobird, Utah.

Kolstad, A. 2001. "Literacy levels and the 80 percent response probability criterion." Chapter 14 in Kirsch, I.S., et al., *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey*, Washington, DC. NCES 2001–457.

Kolstad, A., and D. Wiley. 2001. "On the proficiency penalty required by arbitrary values of the response probability convention used in reporting results from IRT-based scales." Paper presented to the American Educational Research Association, Seattle, Washington.

Kolstad, A., J. Cohen, S. Baldi, T. Chan, E. DeFur, and J. Angeles. 1998. "The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?" NCES Working Paper 2001–20. Washington, DC: National Center for Education Statistics.

Lazer, S., J. Mazzeo, and A. Weiss, with J. Campbell, L. Casalaina, N. Horkay, B. Kaplan, and Rogers, A. 2001. "Final report on enhanced achievement level reporting and scale-anchoring activities" Unpublished report prepared on behalf of the National Assessment Governing Board.

Linn, R.L., and S.B. Dunbar. 1992. "Issues in the design and reporting of the National Assessment of Educational Progress." *Journal of Educational Measurement*. 29 (Summer): 177–194.

Mislevy, R. 1999. "Comments on response probability conventions: a recap of the psychometric perspective." Paper presented to the NCES Response Probability (RP) Convention Meeting, Washington, DC. April 14, 1999

Mullis, I.V.S., and E.G. Johnson. 1994. "The NAEP scale anchoring for the 1992 mathematics assessment." Appendix I, pages 892–908 in E.G. Johnson and J.E. Carlson (eds.), *The*

NAEP 1992 Technical Report. NCES 94–490. Washington, DC: National Center for Education Statistics.

National Assessment Governing Board. March 1995. *Developing Student Performance Levels for the National Assessment of Student Progress: Policy Statement.* Washington, DC: National Assessment Governing Board.

Reckase, M. 2000. *The Evolution of the NAEP Achievement Levels Setting Process: A Summary of the Research and Development Efforts Conducted by ACT.* ACT, Inc.: Iowa City, IA.

Zwick, R., D. Senturk, J. Wang, and S. C. Loomis. 2001. "An investigation of alternative methods for item mapping in the National Assessment of Educational Progress." *Educational Measurement: Issues and Practice*. 20 (Summer): 15–25.