

Running Head: AHM STANDARD SETTING

Standard Setting Using the Attribute Hierarchy Model

Gregory S. Sadesky

University of Alberta

Matthew M. Gushta

American Institutes for Research

### Setting Standards Using the Attribute Hierarchy Model

Setting standards is inherently pattern classification. From the responses of examinees to test items, a classification decision is made reflecting the performance level to which the examinees belong. The challenge in standard setting, as in all pattern classification, is to minimize classification error.

The most prevalent procedures in standard setting rely on the judgment of experts to establish both *cut scores* and *performance standards* (e.g., Zieky, 2001; Kane, 2001). Cut scores are points on the score scale at or above which examinee performance is judged as qualitatively different from those below the score. Performance standards name the specific knowledge and skills that are mastered at each level. The effectiveness of a cut score in differentiating between different levels of achievement as operationalized by the performance standards depend on three critical assumptions. First, do judges adequately identify the skills and knowledge necessary to solve each item? Second, are judges able to accurately assess the probability that examinees at different performance levels will answer each item correctly, as is required in the most popular of judgmental techniques (e.g., Angoff, 1971; Nedelsky, 1954)? Last, is performance on the test sufficiently described by a unidimensional scale such that it is justifiable to differentiate performance levels using scores in a unidimensional metric, usually total score? Each of these assumptions is open to question.

First, the description of the performance standards and the matching up of these standards with test items is an often-neglected step in the standard setting process (Kane, 2001; Sadesky, 2004). As asserted by Kane (2001), beyond a performance label, judges are often not provided with a description of the specific skills and knowledge that define

each performance category. Furthermore, when such standards exist, judges interpret them after the test construction process. Since items are constructed with particular specifications in mind, at best the post hoc standard setting panel could reconstruct the intentions of the test developer and, at worst fail to identify those intentions. Second, Zieky (2001) argues that many of the criticisms of the popular Angoff procedure centre on the difficult, if not impossible task of estimating the probability that a so-called minimally competent examinee will answer a given item correctly. He states that the National Academy of Education (1993) recommended the complete cessation of the use of the Angoff and other item judgment procedures for this precise reason. Third, the unidimensionality of a test is untenable for all but the most monolithic of tests; extra dimensionality can be introduced by many common features of tests including the inclusion of different item formats (e.g., Sireci, 1995; Bolt, 2000), different passages on which items are based, (e.g., Stout et al., 1996), and abstractness of the problem presented by items (e.g., Bolt, 2000). Thus, cut scores in a unidimensional total score metric may not effectively differentiate between true levels of achievement.

In response to these concerns, several researchers have sought to assign examinees to performance categories considering not only their total scores, but also the specific items answered correctly and incorrectly (e.g., Sireci, 1995, Sireci, Robin, & Patelis, 1999; Brown, 2000). Sireci (1995) used cluster analysis to determine whether statistically identified groupings of examinees could be related to the performance categories, and Brown (2000) used latent class analysis to determine if latent distributions of examinees' responses could be used to inform the performance levels to which they belonged. Considering performance on individual items for the purpose of standard

setting is justifiable for at least two reasons. First, the relationship between performance standards and examinee performance is manifest at the level of item responses since the acquisition of a particular skill will only predict performance on items requiring that skill. Second, more than a simple additive measure can determine the attainment of a particular performance level. Instead, the specific pattern of correct and incorrect responses will differentiate levels of performance.

One significant shortcoming of both Sireci's and Brown's work in this area is the lack of consideration of performance standards in the identification of ability categories. Since both the cluster analytic and latent class methodologies employed by this research are driven entirely by characteristics of the data, no direct consideration of the skills and knowledge underlying the items was made. What appears to be needed to address these shortcomings is a method that will retain the advantages of an item level analysis such as those performed by Sireci and Brown but that also considers the relationship of each item to the performance standards. Such a method using the Attribute Hierarchy Model (AHM, Leighton, Gierl, & Hunka, in press), a variation on Tatsuoka's (1983, 1990, 1995) rule space model is described next.

#### The Attribute Hierarchy Model

The AHM was designed to create a description of examinee performance on a test in terms of the cognitive attributes mastered in a given domain. The model works by first identifying all possible states of mastery based on a pre-specified performance model in the domain. Each item on the test is indexed by the attributes that are needed to solve the items and subsequently, vectors are identified that define the expected response patterns for all possible knowledge states in the test domain. The mapping of an examinee's

performance to a specific state of mastery is accomplished based on the likelihood that an observed response pattern is a random deviation of each of the expected response patterns predicted by the performance model.

The relevance of the AHM to standard setting follows from its central characteristic, the ability to place examinees into performance categories. In the typical application of the AHM, the idealized response vectors represent the mastery / non-mastery of discrete cognitive attributes. In the standard setting context, vectors could be used that reflect the performance standards of achievement categories. In this case, the vectors would represent ideal examinee responding as specified by the standards at each performance level. An entire such response pattern for a given level of achievement would thus index all items that should be answered correctly for an examinee at that level.

#### How does the AHM classify examinees?

The AHM assigns individual examinees to performance categories by the *likelihood* that an individual response pattern is a random variation of an expected response pattern. These likelihoods are calculated using item response theory (IRT) item and ability parameter estimates. Item and ability parameters from all examinees are estimated using a 2 parameter logistic IRT model. In addition, the ability ( $\theta$ ) of hypothetical examinees that exhibit each expected response pattern is also estimated. To determine the performance level to which an individual examinee belongs, all deviations of an observed response pattern from each expected pattern are identified. The probability of each deviation from each expected response is calculated based on the 2PL IRT model. The likelihood of a given response pattern originating from one of the

expected response patterns is then determined by the product of the probabilities of all deviations. The performance level for which the likelihood of the expected response pattern is the highest is assigned to the examinee.

Table 1.

Item and Ability Parameters for Hypothetical Example

	Item			Ability level ( $\theta$ )
	1	2	3	
a-parameter	1.0	1.0	1.0	
b-parameter	-1.0	0.0	1.0	
Expected Response 1	1	0	0	-1.0
Expected Response 2	1	1	1	1.0
Observed Response	1	1	0	

To see this more clearly, consider the following hypothetical example. Imagine there is a 3-item test for which two performance levels exist. The item and ability parameters for each item and level are presented in Table 1. Further imagine that an observed response pattern was (1, 1, 0). Given the expected response patterns corresponding to the two levels, which performance level would the AHM assign to the examinee in question? Note first that in comparison with the observed response pattern, there is a deviation of one item response from each expected pattern. The likelihood of the deviation given each expected response pattern determines to which performance level the examinee is assigned. For the lower-achieving level, the deviation is for item 2, and it is of the form  $0 \rightarrow 1$  such that the expected response is 0 and the observed response is 1. What is the likelihood of this deviation? Using IRT, the likelihood of an examinee

of ability  $\theta = -1.0$  correctly answering item 2 is calculated to be 0.27. In contrast, for the higher-achieving level, the deviation is for item 3, of the form  $1 \rightarrow 0$ . Using the same procedure, this likelihood is determined to be 0.50. Given that the likelihood of this latter deviation is higher, the examinee is assigned to the higher performance level.

In the present study, the above methodology will be used to assign individual examinees to performance levels on the 1997 administration of the mathematics portion of the School Achievement Indicators Program (SAIP). This research is intended as a test of the feasibility and interpretability of the AHM as applied to the standard setting context. The performance of this method will be assessed based on an analysis of the relationship of the classification to the underlying item responses and their connection to the performance standards.

### Method

The data used for analysis by the AHM were a subset of examinee data from the 1997 administration of the SAIP test of mathematics. This test comprises 125 items, 25 at each of level 1 through 5. Data from examinees that completed the full set of items determined the subset. Since SAIP is a two-stage test, the majority of examinees did not complete the entire set of test items, though in the case where students were routed to the least difficult subtest, they were not explicitly discouraged from attempting all items. Thus, the data for the present research comprise examinees routed to this subtest who then completed all items. There were 3104 such examinees from which examinees were randomly chosen for analysis by the AHM. The AHM was realized as a program written by Steve Hunka using Mathematica 4.1. Consistent with the performance level taxonomy for the SAIP, 6 levels (0 through 5) and associated expected response patterns were

defined for the purposes of the AHM. All examinees were assigned to levels of achievement based upon the highest value for the likelihood of the deviation of their observed responses from the expected responses associated with each level.<sup>1</sup>

### Results

Table 2.

Total Score Means and Standard Deviations by Performance Level

	Performance Level		
	1	2	3
Mean	38.98	58.78	74.53
Standard Deviation	6.63	7.90	4.11
Number of Examinees	40	40	40

Note: Total score = 125

Since the data for the present sample comprised only those examinees routed to less difficult subtest, very few examinees attained above level 3. Therefore, the present analysis will focus on the characteristics of the responses of 40 randomly chosen examinees that were classified to levels 1 through 3. The means and standard deviations of the number of items answered correctly for each group is displayed in Table 2.

Interestingly, there was considerable overlap of the total scores achieved by examinees assigned to adjacent performance levels. That is, there are a number of examinees at the

<sup>1</sup> Significant differences exist between the criteria for classification designed for SAIP and that of the AHM. In order for an examinee to attain a level under the SAIP criteria, a minimum number of items had to be answered correctly within each content area, within each level. For example, to attain level 2, examinees must meet all of the following criteria, a) 6/11 number and operations, b) 1/3 algebra, c) 5/8 measurement and geometry, d) 1/3 data management and statistics, and e) 15/25 items total must be answered correctly. The highest level for which an examinee attained all criteria was designated as his or her level of achievement.

boundary of classes 1 and 2 and of classes 2 and 3 for which the total number correct for the examinee at the higher level was less than the total for an examinee classified to the lower level. What characterizes these examinees? Table 3 lists the mean number of items targeted at each performance level that were answered correctly for examinees falling at the classification boundary. The pattern of responses for examinees classified to the lower of the two levels is consistent. Though the overall number of items answered correctly is greater than those classified to the higher level, they correctly answered *fewer* of the items targeted at the higher level. For the boundary between levels 1 and 2 the average increase in the number of such items answered correctly is 2.4, while for the level 2 / 3 boundary, the increase was 1.6. It is interesting to note that some examinees classified to the lower achieving group answered more of the level 4 and 5 items correctly but given that the likelihood based on the 2PL model of answering these items correctly is low for this group, they had little impact on the overall classification. Table 3.

Mean Number of Items Answered Correctly by Level for Groups Close to Classification Boundary (n=5 per row)

	Level of Item					Mean Total
	1	2	3	4	5	
Class 1 – Below Boundary	20.6	10.6	10.2	4.6	3.6	49.6
Class 2 – Above Boundary	18.4	13.0	8.0	3.6	3.2	46.2
Class 2 – Below Boundary	23.2	21.4	14.6	7.8	3.2	70.2
Class 3 – Above Boundary	22.0	21.0	16.2	5.2	3.0	67.4

## Discussion

Demonstrated above, the AHM produces an appropriate and interpretable classification of examinees into performance categories. This classification method has several desirable properties that can address the shortcomings of current standard setting methods based on the judgments of post hoc panels.

First, the connection between items and performance standards is made explicit. This results because the items targeted at each level defined the expected response patterns. The classification of an examinee to a level occurs because the items that reflect the performance standards at a given level are likely to be answered correctly.<sup>2</sup>

Second, the need for post hoc standard setting panels is reduced because the connection between the item and performance level has already been made, possibly at the test development stage. Since convening these panels demands significant resources, using the AHM as a standard setting method could represent a considerable increase in efficiency.

Last, cut scores using the AHM are not simple discriminants made in a unidimensional total score metric. Instead, they reflect the relationships of specific items to specific performance levels and thus can be used to make more precise discriminations between examinees with similar total scores. In particular, those who have correctly answered items that are directly related to the performance standards are more likely to be assigned to the level associated with those standards as compared with an examinee whose correct answers do not consistently relate to these standards.

---

<sup>2</sup> Notably, the central task in the modified Angoff procedure, the estimation of the class conditional item probabilities, can be derived directly using the IRT model.

### Conclusion

As a result of the above demonstration, the AHM can be seen as a promising technique for standard setting. The specific advantages of the technique include the direct connection of classification decisions to the attainment of the performance standards, the lack of reliance on post hoc panels to evaluate the probability that examinees at given levels will answer items correctly, and the ability to make categorization decisions without being locked into a unidimensional total score framework. Further research on the AHM in standard setting could focus on a direct comparison between the AHM and judgmental methods on specific operational tests.

## References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2<sup>nd</sup> ed.). (pp. 508-600). Washington, DC: American Council on Education.
- Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, 37, 307-327.
- Brown, R. S. (2000). *Using latent class analysis to set academic performance standards*. Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53 – 88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Leighton, J. P, Gierl, M. J., & Hunka, S. (2002). *The attribute hierarch model: An approach for integrating cognitive theory with assessment practice*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.
- National Academy of Education. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education
- Nedelsky, L. (1954). Absolute grading for objective tests. *Educational and Psychological Measurement*, 14, 3-19.

- Sadesky, G. S. (2004). *The use of cluster analysis in standard setting*. Manuscript submitted for publication.
- Sireci, S. G. (1995, August). *Using cluster analysis to solve the problem of standard setting*. Paper presented at the meeting of the American Psychological Association, New York.
- Sireci, S. G., Robin, F. R., & Patelis, T. (1999). Using Cluster Analysis to Facilitate Standard Setting. *Applied Measurement in Education*, 12(3), 301-323.
- Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20(4), 331-354.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.
- Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.