DOCUMENT RESUME

ED 379 335 TM 022 714

AUTHOR Hansche, Linda
TITLE Technical Issues in Performance Assessment: Setting Performance Standards.
PUB DATE Jun 94
NOTE 17p.; Paper presented at the Annual Meeting of the National Conference on Large Scale Assessment (Albuquerque, NM, June 14-17, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Decision Making; *Educational Assessment; Educational Policy; Elementary Secondary Education; *Interrater Reliability; Multiple Choice Tests; Pilot Projects; *Scoring; *Standards; *Test Construction; Test Items
IDENTIFIERS Dominant Profile Procedure; Judgmental Policy Capturing; *Performance Based Evaluation; *Standard Setting

ABSTRACT

Setting standards on performance measures is discussed in the context of the State Collaborative on Assessment and Student Standards (SCASS) initiative supported by the Council of Chief State School Offices. The usual item-based methods for standard setting, the methods developed by Nedelsky (1954), Angoff (1971), and Ebel (1972), were developed for use with large numbers of multiple choice items of a unidimensional nature with item scores contributing to a summative scale. They have little or no value when applied to performance items, where scores typically do not measure a single construct, and there are often multiple scores for each exercise. One approach that is being considered for performance standard setting is judgmental policy capturing, a general procedure designed to describe statistically the unique information processing strategies of individual raters. It usually includes the derivation of one or more summary policies that characterize all raters. Pilot tests have supported the utility of this approach. Another approach is that of the dominant profile procedure, in which a standard setting panel first states and then redefines policies as they attempt consensus about successful scores. SCASS efforts will consider both these approaches and their generalizability. A handout is attached. (Contains 3 references.) (SLD)

Technical Issues in Performance Assessment:

Setting Performance Standards

Dr. Linda Hansche, Georgia State University

Presented at the annual meeting of the National Conference on Large Scale Assessment, Albuquerque, NM, June 14-17, 1994.

I have been invited to speak to you today about setting standards on performance measures as part of the State Collaborative on Assessment and Student Standards (SCASS) initiative supported by the Council of Chief State School Offices (CCSSO).

Let me begin by saying that my experience with standard setting has been considerable from a user's perspective. I have been heavily involved in developing high stakes tests for both students and teachers for the past 15 years. Conducting standard settings, at least for multiple choice items has become almost routine. Recently in the past year, I have been involved in the CCSSO SCASS project investigating technical issues related to developing and implementing performance assessments. One of my interests has been in how to set standards on performance items as well as how to integrate that process with standards for traditional multiple choice items.

Georgia is presently enjoying what I feel is an enviable position. We have a state mandated curriculum, the Quality Core Curriculum, which the Department of Education is required to assess for accountability purposes. As a contractor to the state we have developed item banks which are administered on a matrix sampling basis to students in grades 3, 5, and 8. The only requirement tied to the item bank developed and used for assessing the curriculum is that it be one-third higher order thinking skills. Though the mandate has been in place since the

thinking skills. Though the mandate has been in place since the late 1980's, the program is quite in vogue these days. As Tony Nitko described in his presidential address to NCME this past April, curriculum driven assessment is the assessment of the present. Actually called the Curriculum-Based Assessment Program, Georgia's student assessment program is right in line with various educational reform issues which are tying assessment and curriculum tightly together. Furthermore, Georgia does not have a mandate for performance or alternative assessments -- yet. We are able to experiment with alternative item types and begin to formulate policy recommendations before someone, i.e. the state legislature or state board of education, requires us to provide such assessments -- thus our enviable position.

As part of our initiative to understand the shifting assessment landscape, to evaluate existing research, and to embark on our own research agenda, I have been representing Georgia as a member of the SCASS group pursuing technical issues surrounding performance assessments. As part of my personal interest, I helped to formulate a team to investigate setting standards for performance items. Other states on this team include Virginia, Louisiana, and Utah. We engaged Dick Jaeger to produce a paper including a through review of the literature, a statement of the problem, and a proposed research agenda relative to setting standards for performance assessments. A proposal is in final draft form and will become the basis for an RFP for which CCSSO will seek funding. My handout is excerpted from that

proposal. At this point in time, we are hopeful that monies will be found to address this needful issue.

That catches you up on who I am and why I'm interested in standard setting and am standing in for Dick. Now, let me tell you as best I can about the status of the proposed study for setting standards on student performance assessments. Let me begin by providing some background on why current methods of setting standards do not apply to performance assessments and what alternative methods have been piloted.

In the past, there have been three item-based methods for standard setting. Developed by Nedelsky (1954), Angoff (1971), and Ebel (1972), each of the methods requires that judges independently conceptualize a minimally qualified examinee, and then estimate the probability that such an examinee would make appropriate judgements concerning each item on a test. The methods differ in what they describe as an appropriate judgement.

Nedelsky's (1954) method requires that judges independently predict the number of response options on each multiple-choice test item that a "minimally acceptable" examinee should be able to eliminate as being clearly incorrect.

The Angoff (1971) method requires a judge to estimate the probability that a "minimally acceptable" person would answer each item correctly.

Ebel's (1972) method involves three steps. First each judge classifies the items on a test using a difficulty dimension, for which there are three categories (easy, medium, and hard). Next the judge classifies the items using an importance dimension for which there are four categories (essential, important, acceptable and questionable). The judges then estimate the proportion of items in each category that a "minimally acceptable" person would answer correctly.

With all three methods, an overall test standard is set by computing averages of all sampled judges. When applied to a statewide assessment, a "minimally acceptable" examinee would be defined as one with performance that in the judges' view was just sufficient to warrant the goal.

Numerous variations of these methods abound. One much used modification to the Angoff procedure was instituted by Dick Jaeger himself. Jaeger successfully modified Angoff's method to provide multiple opportunities for judges to consider and reconsider their judgements. As modified by Jaeger, between iterations of the judgement process judges are provided with empirical difficulty information for each item, and the distribution of recommendations by other judges on the panel. A discussion is held among the judges before they are given the opportunity to reconsider their initial judgements.

Each of the three methods described -- Ebel, Angoff, and Nedelsky -- were developed for use with large numbers of multiple choice items of a unidimensional nature with item scores

contributing to a summative scale; as such, they have little or no value when applied to performance items. Setting standards by these methods requires that a panel of experts judge the difficulty of individual test items. These judgements are then summed across the items for each judge and then averaged across all judges.

This fundamental strategy is unacceptable for performance items for several reasons. First, performance assessments are typically scored at the exercise level with exercise scores typically not measuring a unitary construct. Second, many performance assessment exercises are scored polytomously rather that dichotomously. Third, performance assessments often yield multiple scores for each exercise, producing complex sets of data along imperfectly correlated dimensions. And finally, since performance assessments are usually complex and relatively lengthy to administer in terms of time, typically only a few exercises are usually administered.

The measurement community seems to be in agreement that new methods for standard setting must be created which account for the complex nature of performance assessments. The new standard setting procedure or procedures must acknowledge and incorporate differences in examinee performances on tasks or exercises of varying importance relative to overall ability. And it almost goes without saying that any new method must be defensible when subjected to strict psychometric scrutiny.

In response to the need for alternative methods for setting standards for performance items, Jaeger and his colleagues have piloted two methods in conjunction with work undertaken for the National Board for Professional Teaching Standards. (A third method, an Extended Angoff method, was also piloted by Ron Hambleton but is not reported here.) The first of the methods is a procedure called Judgmental Policy Capturing. A full report of this study was presented at AERA this past April and the paper can be requested from Dr. Jaeger at UNC-G.

Judgmental policy capturing methods have been applied in the context of industrial and organizational psychology in areas such as clinical judgements, personnel selection and promotion decision, human learning, and interpersonal conflict. Judgmental policy capturing as a general procedure is designed to describe statistically the unique information processing strategies of individual raters and usually includes the derivation of one or more "summary" policies that characterizes all raters. The first step in this procedure is to present expert raters with a series of performance profiles consisting of different dimensions of performance. The second step is to request raters to review each profile and then assign an overall rating that best represents or summarizes the information contained in the profile. The third step is to calculate the extent to which an individual rater's overall ratings are predictable and to compute the relative importance of each single dimension in determining overall ratings. The resulting equations define the "captured policy"

for each individual rater. This captured policy is taken to represent an explicit objective description of the way in which raters combine and weight dimensional information in arriving at overall ratings.

Judgmental policy capturing was piloted in conjunction with the National Board for Professional Teaching Standards' quest to set performance standards for early adolescence English language arts teachers. Five exercises were selected from the National Board's Early Adolescence English Language Arts assessment package. Twelve expert teachers in the field composed the standard-setting panel. Panel members were carefully trained in the use and scoring of the five selected exercises. The policy capturing was conducted in two stages. First, individual exercises were judged and five equations were produced. Then panelists used all five exercises to predict overall performance scores across exercises.

The results of the study were encouraging. Judges were able to be adequately trained in the procedure; the panelists were able to judge approximately 200 profiles in a reasonable amount of time without undue fatigue; and most panelists were highly consistent in their responses. Thus the method would appear to be reasonable for use with performance assessments. However, as with any procedure, there are some limitations. The procedure does not provide panelists with information related to the implications of their recommendations. Failure to provide panelists with this information also fails to allow for

enlightened discussions and the opportunity to reconsider initial judgements in light of some type of impact data. The study concluded with the stated need for further research into judgmental policy capturing by incorporating an iterative process.

On the same research agenda to develop sound methods for setting performance standards for the National Board for Professional Teaching Standards, the dominant profile procedure was also pilot tested. In a procedure called a multi-stage dominate profile, a standard setting panel engaged in a process of first stating and then refining their policies as they attempted consensus about what configurations of scores on a multi-exercise assessment would indicate success for certification. A report of this research was presented at AERA by Sarah Putnam and copy of the paper can be obtained by contacting her at University of North Carolina, Greensboro.

There were three stages for the dominant profile procedure reported in the study. The first stage was one of policy creation. Expert panelists who were familiar with the assessment exercises and scoring procedures were given ten blank profiles on which to create their "bottom-line" or "just barely certifiable" candidate. Profiles were completed along with a brief description of the policy that guided the panelists' decision making. Each profile was analyzed to determine content and to investigate panelists' areas of agreement and disagreement.

In stage two of the process, one week later, panelists received feedback on their own standard setting recommendations including graphs of their own profiles along with similar data for the other panelists. Panelists also received a packet of new profiles they were to use to make further standard setting recommendations. Panelists were then asked to study the data and to complete their recommendations for the new profiles.

For stage three, "challenge" profiles were developed by the researchers based on the panelists' proposed policies. Most profiles contained a score below a panelist-specified lower bound on a dimension which would then be offset by a higher score on the same dimension in another exercise. Panelists were instructed to respond to each profile individually either "yes" for pass or "no" for fail.

Analyses of stage three data indicated that only 65% of the profile classifications could be predicted accurately. The researchers were not hopeful that a single logistic regression model could be used. They also stated that the model's failure to account for the conjunctive judgement policies was a severe limitation. Unless panelists are able to agree on a single "bottom-line" policy, the procedure is at an impasse, and an alternative analytic model must be sought.

In recommending further development for a multi-stage dominant profile method, the researchers felt that panelists must have the opportunity to examine their explicit responses and then to alter those policies if they so choose. Some form of

iteration appears to be essential in setting standards using this or any other procedure.

In summing up the studies conducted under the auspices of the National Board, it was determined by the researchers that the dominant profile method used in conjunction with the judgmental policy capturing provided the most fruitful line for further research. By engaging panelists first in a judgmental policy capturing activity followed by thoughtful reflection and participation in the multi-stage dominant profile method including iterations, it was felt that a strong method for setting performance standards could be developed.

Which brings us back around to the current topic -- the SCASS project to research technical issues related to setting standards for large-scale performance assessments. As I stated earlier, Dick Jaeger has submitted a proposal for researching further a method for setting performance standards. The purpose of the research is to develop new methods of standard-setting for applications to student performance data, for formulation of criteria and strategies for evaluating the validity and stability of resulting student classifications, and for testing these procedures using the statewide performance assessment data of several states.

An original proposal designed one study to investigate a dichotomous standard for individual student data and a second study to investigate setting polytomous standards on aggregated data. After our last SCASS meeting in March, it was decided that

the proposal is to be modified to focus on the development of a procedure or procedures for setting polytomous standards for large sets of individual student performance data. One reason for this decision, in addition to limited time and money resources, is that any procedures found utile for these data would most likely be generalizable to other situations such as aggregated data sets used for program evaluation or accountability purposes. The research will investigate and perhaps actually incorporate both the judgmental policy capturing method and the multi-stage dominant profile procedure. We are hopeful that a funding agent has been identified and are awaiting agreements on final changes in the proposal before undertaking this study.

Stay tuned for the next breaking news flash in the mystical world of performance standard setting. . .

Technical Issues in Performance Assessment:

Setting Performance Standards

Dr. Linda Hansche, Georgia State University

Handout prepared for presentation at the annual meeting of the National Conference on Large Scale Assessment, Albuquerque, NM, June 14-17, 1994.

Excerpted from Methods for Setting Standards on Performance Assessments in Statewide Assessment Context: A Proposal submitted by Richard M. Jaeger to the Council of Chief State School Officers State Collaborative on Assessment and Student Standards by request of the Technical Guidelines for Performance Assessments planning team.

"Policy capturing (see Jaeger, Plake, & Hambleton, 1993; Jaeger, 1994, April) is a judgmental process that attempts to elicit and characterize the decision strategies employed by expert judges when they evaluate profiles of examinees' performance on multiple exercises and reach a summative decision concerning the overall quality of the examinees performances. In this process, expert judges respond independently to a large number of simulated profiles of examinees' performances on the elements of an assessment package, indicating for each stimulus profile, their evaluation of the overall performance of the examinee. These profile-response pairs are then used to "capture a judge's policy" in evaluating overall examinee performance. Various analytic procedures are applied to the data contained in the profile-response pairs to determine the relative weights judges apply to elements of examinee performance in reaching an overall evaluation, and the range of performance profiles associated with a recommendation to "pass" an examinee. The method produces a mathematical representation of the judgment

policy of each member of a judgment panel, an estimate of the decision consistency of each panel member, and information on the inter-judge consistencies of panel members.

The multi-stage dominant profile procedure presents judgement panelists with opportunities to specify the profiles to examinee performances that, in their judgement, define the lower boundary of passing performance; to evaluate their recommendations in the context of those provided by their fellow panelists; to discuss their recommendations with their fellow panelists, and to reconsider their initial recommendations in light of discussion and information on the consequences of their recommendations. This procedure has been found by Pence, Putnam, and Jaeger (1994) to be effective in fostering convergence of judgements among panelists if it is used as an adjunct to judgmental policy capturing. It appears, as well, to produce performance standards that are richly grounded in the complexities of typical performance assessments and authentically reflect the analytic reasoning of standard-setting panelists.

References

Jaeger, R.M., Plake, B.S. & Hambleton, R.K. (1993, April). Integrating multi-dimensional performances and setting performance standards. Presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA, April 13-15, 1993.

Jaeger, R.M. (1994, April) Setting performance standards through two-stage judgmental policy capturing. Presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA, April 3-8, 1994.

Pence, P., Putman, S. & Jaeger, R.M. (1994, April). A multi-stage dominant profile procedure for setting standards on performance assessments. Presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA, April 3-8, 1994.