DOCUMENT RESUME

ED 413 362 TM 027 701

AUTHOR McGinty, Dixie; Neel, John H.; Hsu, Yu-Sheng
TITLE Interjudge Variability and Intrajudge Consistency Using the Cognitive Components Model for Standard Setting.
PUB DATE 1996-00-00
NOTE 30p.; Paper presented at the Annual Meetings of the Georgia Educational Research Association (Atlanta, GA, 1996) and the American Educational Research Association (Chicago, IL, March 24-28, 1997). For a related study, see TM 027 700.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cognitive Processes; Criterion Referenced Tests; Elementary Education; Evaluation Methods; Grade 3; *Interrater Reliability; *Judges; *Mathematics Tests; Primary Education; Scoring; Standards; *Test Items
IDENTIFIERS Angoff Methods; *Cognitive Component Analysis; *Standard Setting

ABSTRACT

The cognitive components standard setting method, recently introduced by D. McGinty and J. Neel (1996), asks judges to specify minimum levels of performance not for the test items, but for smaller portions of items, the component skills and concepts required to answer each item correctly. Items are decomposed into these components before judges convene. A preliminary study supported the usefulness of the approach and suggested that it was equal to the Angoff method in most respects and resulted in lower interjudge variability. In this study, the original study was replicated with a different set of judges. Twelve judges, all third- and fourth-grade teachers, set standards for a Georgia criterion-referenced mathematics test for third grade. Forty-five test items were used. Each judge provided ratings using both the Angoff method and the cognitive components method. Results are encouraging with regard to the potential of the cognitive components model as an alternative to standard setting. As in the initial study, the cognitive components model resulted in lower variability among judges at all levels of the process. This may suggest that a substantial proportion of the disagreement among judges using the Angoff method is due to judges' differing abilities to perceive the important characteristics of items. (Contains 6 tables and 21 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running Head: COGNITIVE COMPONENTS STANDARD SETTING MODEL

ED 413 362

Interjudge Variability and Intrajudge Consistency Using the Cognitive Components Model for Standard Setting

Dixie McGinty

John H. Neel

Yu-Sheng Hsu

Georgia State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Dixie McGinty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM027761

Paper presented at the 1996 annual meeting of the Georgia Educational Research Association, Atlanta, and the 1997 annual meeting of the American Educational Research Association, Chicago

BEST COPY AVAILABLE

Interjudge Variability and Intrajudge Consistency Using the Cognitive Components Model for Standard Setting

Despite several decades of research, setting standards for minimal competency tests remains problematic. The popular judgmental methods proposed by Angoff (1971), Ebel (1972), Jaeger (1989) and Nedelsky (1954) have much to recommend them, and admirable efforts continue to be made to refine these processes. Nevertheless, standard-setting procedures continue to be fraught with difficulties that make them vulnerable to harsh criticism. One of the most salient problems is that the recommendations of the judges are often substantially more variable than might be hoped, which reduces our confidence in the standard that has been set.

Variability in judges' recommendations for a set of items can result from either (1) differing opinions about what should be required of examinees or (2) differing perceptions of the test items and the cognitive demands they pose. We suggest that the first type of variability is to be expected; we expect individuals to differ in their opinions and in the value judgments they make. The second type of variability, on the other hand, is potentially more threatening; it results from judges' varying abilities to perceive correctly the important features of test items.

We believe that the judges' task in standard setting can, and should be, made easier, especially with regard to the perception of items. Judges, even those who are teachers, typically have limited experience with actual test items, and many lack training in cognitive psychology. It seems unrealistic to expect them to become skillful in assessing the difficulty of items after just a few hours of training. Further, the time allowed for standard setting may not be sufficient for judges to thoroughly analyze each item and identify the skills it demands. We thus believe that a new method is needed, one that takes some of the guesswork out of the prediction of item difficulty.

One such method, the cognitive components method, was recently introduced by McGinty and Neel (1996). Using this method, judges specify minimal levels of performance not for items themselves, but rather for smaller pieces of items, i.e., for the component skills and concepts that are required to answer each item correctly. In a preliminary empirical investigation (McGinty & Neel, 1996), the cognitive components model resulted in lower interjudge variability than did the Angoff method, while equaling the Angoff method in other aspects (for example, correlations between MPLs and empirical item p-values were high for both methods, supporting the validity of the process). The results of this preliminary study suggest to us that the cognitive components model is indeed worthy of further research. Replications of the initial study, using different judges, are needed. Also, other aspects of the new model should be examined to answer questions not addressed by the previous study. One such aspect is intrajudge consistency.

The current study addressed two questions:

(1) Would a replication of the original study using a different set of judges yield similar findings with regard to interjudge variability?

(2) Does the cognitive components model offer any advantages over the Angoff model with regard to intrajudge consistency?

The Cognitive Components Model

The cognitive components model (McGinty & Neel, 1996) is an alternative model for judgmental standard setting. Like the Angoff procedure, the cognitive components method generates a minimum pass level (MPL) for each item, which is equivalent to the estimated probability of a correct response by a hypothetical "minimally competent examinee." In both approaches, the item MPLs are summed for each judge, and the sums are then averaged across judges to yield the standard, or minimum passing score for the test. What makes the cognitive components method very different from the Angoff method is the way in which MPLs are obtained. Using the Angoff method, each judge examines each item directly and simply estimates the probability that a minimally competent examinee will respond correctly. With the cognitive components method, in contrast, judges assign probabilities not to entire items, but to more specific skills, concepts, or subtasks that are presumed to be necessary for a correct response. The relevant probabilities for each item are then multiplied to arrive at the probability of a correct response to the item as a whole, i.e., the MPL. The computation of the MPL is described in greater detail in a later section.

Before judges convene, items are decomposed a priori into nonoverlapping "cognitive components," which may be thought of as specific skills, subtasks, or pieces of knowledge that are assumed to be required for a correct response to an item. Consider, for example, the following estimation item, which is similar to one item on the test used in this study (assume that the response options are all multiples of 100):

516 + 193 + 232 is about ________.

One way to decompose this item is to postulate that, in order to respond correctly, an examinee must (a) recognize "about" as a prompt for rounding or estimation, (b) round three-digit numbers to the nearest hundred, (c) recognize "+" as a prompt for addition, (d) line up numbers vertically for addition, and (e) apply basic addition facts.

The decomposition of all items on a test results in a larger set of cognitive components that are represented in various combinations by the individual items. Ideally, no component is unique to a particular item, though in reality this probably will not hold true unless the components were identified during the test development process.

When judges convene, they are presented not with actual test items, but with brief statements or descriptions of cognitive components. For each, the judges are asked to complete a statement of the type, "In order to be considered competent (or, to be promoted, etc.) an examinee must be able to apply this skill correctly at least _____% of the time." In other words, judges specify the minimum ratio of the number of correct applications of the specific component to the number of situations that require it (note that this is not equivalent to the proportion of items requiring the skill that should be answered correctly). This value is called the minimum success rate (MSR) for the cognitive component. It is equivalent to the probability that a minimally competent examinee will apply the cognitive component successfully.

Decomposition of intellectual tasks into component subtasks is not at all without precedent in cognitive psychology. Donders (1969), S. Sternberg (1969), and R. J. Sternberg (1977, 1978, 1979, 1983), developed experimental methods for isolating discrete components of tasks and found empirical evidence for subtask independence. These and other such research efforts, reviewed by Sternberg (1979) and Pellegrino and Glaser (1979) indicate that component processes have been used successfully to model examinee performance on a number of test item types, including verbal and geometric analogies, spatial transformation items, and others. The theoretical cognitive-psychometric work of Fischer (1973) and Embretson (1983, 1984, 1985b) involves mathematical modeling of test item difficulty based on item characteristics or subtasks that are identified a priori. There is thus support in the literature for the idea that test items can be broken down into parts and that item difficulty can be modeled using these parts.

Underlying the cognitive components standard-setting model is the assumption that the probability of a correct response to an item can be modeled as the product of the probabilities of successful application of independent component skills or concepts. That is, the probability that the total item i can be correctly answered by an examinee j is given by

$$P(X_{ij} = 1) = \prod_{k} P(X_{ijk} = 1),$$

where $P(X_{ij} = 1)$ is the probability that total item i is correctly answered by examinee j, and $P(X_{ijk} = 1)$ is the probability that examinee j successfully applies component skill k. The same model has provided the basis for cognitive-psychometric work by other researchers (e.g., Embretson, 1985b; Whitely, 1980).

In standard setting, specifying a minimum pass level (MPL) for an item is equivalent to estimating the correct-response probability for a given examinee, specifically, that examinee whose ability level defines the border between competence and lack of competence. The MPL is thus an estimate of $P(X_{ijT} = 1)$ for the hypothetical examinee j whose ability level defines minimal competency. Similarly, if a cognitive components model for standard setting is assumed, the minimum success rate (MSR) specified for a single cognitive component k is an estimate of $P(X_{ijk} = 1)$.

Using the Angoff method, item MPLs are estimated directly. With the cognitive components approach, in contrast, an item MPL is defined as the product of the MSRs for its component parts. As an illustration, consider the hypothetical rounding item presented earlier in this chapter. The four cognitive components postulated for this item were components actually used in this research (the item, however, was not), and their average MSRs across judges in the first of the two studies to be described here were .775, .667, .883, and .973. The synthetic MPL for this item would thus have been the product of these, which is .4441.

Methods

Instrument

The instrument was the mathematics subtest of the Georgia Criterion-Referenced Test for third grade. The test, which was used to determine whether students were eligible for promotion to the fourth grade, was administered in the spring of 1991 to a statewide population of approximately 90,000 third graders. The mathematics subtest consisted of 85 four-option multiple-choice items. The reported Kuder-Richardson reliability coefficient was .915; the standard error of measurement was 2.892. While most items tested simple computation skills in addition, subtraction, and multiplication, a few other types of items were included, such as items requiring the examinee to select an appropriate unit for a measurement task, or to select the appropriate operation for a word problem. The test consisted largely of rather easy items, with a mean p-value of .86. Fifty-five of the 85 items on the test were used in this research.

Item Decomposition

Before standard-setting data could be collected from judges for this research, it was necessary to decompose each item on the test into cognitive components or subtasks. This was an a priori process. Cognitive components for each item were proposed by the authors together with a mathematics educator who had also been employed by the organization that developed the test. Two third grade teachers, both of whom had served as item reviewers in the development of this test, were also called upon as needed for advice about specific aspects of the process. The set of cognitive components used in this research consisted of the 29 components listed in Table 1. The number of components identified for each item ranged from one to seven.

Data Collection

Standard-setting judgments were collected during the last two weeks of May, 1996 from 12 judges, all of whom were currently teaching third or fourth grade in public schools in Georgia. Both third and fourth grade teachers were deemed competent to make decisions about the test since passing it had previously been a requirement for promotion from third to fourth grade.

Table 1

Cognitive Components Used in this Study

C1. Translate words to numerals.
C2. Choose the correct operation to solve a word problem.
C3. Count objects in a picture.
C4. Understand what is meant by "tens" and "ones" in place value.
C5. Compare two numbers to determine which is greater.
C6. Apply basic addition facts.
C7. Line up amounts of money vertically for computation.
C8. Regroup (in addition).
C9. Recognize "+" as a prompt for addition.
C10. Compare three or more numbers to determine which is greatest.
C11. Read a table.
C12. Know the monetary value of a pictured coin.
C13. In a subtraction word problem, know which number to subtract from which.
C14. Apply basic subtraction facts.
C15. Select an appropriate unit of measure.
C16. Recognize "-" as a prompt for subtraction.
C17. Round three-digit numbers to the nearest hundred.
C18. Line up two- or three-digit numbers vertically for computation.
C19. Compare sizes of pictured objects.
C20. Recognize "about" as a prompt for estimation or rounding.
C21. Know what is meant by perimeter of a figure.
C22. Regroup (in subtraction).
C23. Read a bar graph.
C24. Recognize "÷" as a prompt for division.
C25. Recognize "×" as a prompt for multiplication.
C26. Apply basic multiplication facts.
C27. Recognize ")" as a prompt for division.
C28. Know the monetary value of a coin by its name.
C29. Apply basic division facts.

Six of the judges taught at one school, another 3 taught at a second school, and the last 3 taught at a third school. Compensation for participation was $50.00 for one session of approximately three hours.

To collect the data, simulated standard-setting meetings were conducted with groups of 3 judges at a time. All meetings were held at the judges' schools after school hours. The decision not to have all 12 judges convene on a single occasion was made for two reasons. First, there was no pressing reason, outside of the need for standardization of implementation, why judges needed to be together, since the standard-setting method used was not iterative. Second, greater incentives would probably have been needed in order to ensure the necessary level of participation if all judges had been required to travel to a particular site on the same date.

At the meetings, each judge provided ratings using both the Angoff method and the cognitive components method. In other words, each judge responded both to a set of items and to the set of cognitive components that had been identified for those items. Two of the four groups of judges completed the Angoff task first, followed by the cognitive components task; the other two groups performed the tasks in the reverse order. Assignment of the groups to the Angoff-first or the cognitive-components-first condition was done randomly.

Before each task, judges were given detailed oral instructions about the procedure to be used, along with a handout outlining the most important points. Instructions for the two methods were designed to require approximately the same amount of time. For both methods, judges were encouraged to ask as many questions as necessary to clarify the procedures.

For the Angoff method, each judge received a packet of 55 pages with one photocopied test item at the top of each page. The judge was to complete the following sentence, which appeared below each item: "Out of 100 minimally competent examinees, _____ should answer this item correctly." There was ample space remaining on each page for judges to write comments about the items, and they were encouraged to do so.

For the cognitive components method, each judge received a packet of 29 pages with a statement of a cognitive component printed at the top of each page. Below each cognitive component was the statement, "In order to be promoted to the fourth grade, an examinee should be able to apply this skill or knowledge correctly at least _____ percent of the time." As with the Angoff method, judges were encouraged to write comments about the components if they wished.

Before beginning each task, judges were told that they must work independently, though they were free to ask questions of the experimenter if necessary. They were allowed as much time as they needed to complete the task. When all three judges had finished a task (i.e., the entire set of items or cognitive components, depending on the method), instructions were presented for the other task.

Data Analysis

For each judge, the raw data obtained in the study consisted of MPLs for each item using the Angoff method and MSRs for each cognitive component. Computing the product of the relevant MSRs for each item yielded a synthetic MPL for each item-judge combination. For each of the two methods, a standard was then computed by summing across items the MPLs for each judge, then averaging these sums across judges.

Variability among judges was investigated in a number of ways. First, means and standard deviations of the raw responses (i.e., item MPLs for the Angoff method, component MSRs for the cognitive components method) were examined. The average standard deviation for the MPLs was compared to the average standard deviation for the MSRs in order to determine whether judges tended to agree more about cognitive components than about items, or vice versa. Another very important question was how the variability of the synthetic MPLs yielded by the cognitive components method would compare with that of the Angoff MPLS. To address this, standard deviations across judges were compared item by item.

Interjudge variability was also investigated by examining all possible interjudge correlations within each method and comparing the proportion of significant correlations yielded by cne method to the proportion yielded by the other method. This also made it possible to identify any judge(s) whose responses were not at all correlated with those of the other judges. Intercorrelations of the MSRs among judges were also examined.

Intrajudge consistency is more difficult to investigate, since judges do not ordinarily have the opportunity to rate the same items on multiple occasions. Van der Linden (1982) proposed a measure of intrajudge consistency based on item response theory (IRT). IRT models the probability of success on an item as a function of (a) the ability level of the examinee, and (b) one to three item parameters. Using the three-parameter IRT model, which was used in this study, the probability that an examinee of a given ability will answer an item correctly is given by

$$P_i(X=1|\theta)=c_i+(1-c_i)\frac{1}{1+e^{-a_i(\theta-b_i)}},$$

where a_i is the discrimination parameter for item i, b_i is the item difficulty, c_i is a pseudo-guessing parameter, and θ is the ability of the examinee on the trait being measured.

A cutoff score arrived at through a standard-setting procedure is, in effect, an ability level; specifically, it is the ability level of the hypothetical "minimally competent" student. If item parameter estimates are available, then, we can estimate the probability that the minimally competent examinee will answer each item correctly.

Inconsistency occurs when one judge specifies probabilities of item success that are incompatible with each other or with the ability of that judge's own conception of a minimally competent examinee as represented by the standard resulting from his/her ratings. Suppose, for example, that a given judge's standard was set at a raw score of 45 items, and that this score corresponded to θ = .5 on the ability scale. Suppose also that this judge had assigned an MPL of .75 to a particular item. If that item had parameter values of $\underline{a}$ = 1.25, $\underline{b}$ = .80, and $\underline{c}$ = .20, simple computation using the 3-parameter IRT model shows that the estimated probability of success on this item for an examinee with θ = .50 is .53. The judge has assigned to the item a probability of success that is inconsistent with the borderline student he/she has in mind. In a similar way, a judge may assign several item probabilities that could never belong to the same examinee.

Van der Linden's index of intrajudge consistency is based on a judge's average specification error in rating the items. The specification error is defined as the absolute value of the difference between the MPL a judge specifies for an item and the value of the item response function evaluated at the value of θ that corresponds to the judge's overall standard for the test. The judge's average specification error is then transformed to account for the fact that the range of values it may take on depends on how extreme the standard is. In other words, if the IRT model yields a probability that is very close to zero or to one, the maximum possible specification error is larger than it would be if that probability were closer to .50.

For two reasons, Van der Linden's index seems to be the best available indicator of the consistency of judges' ratings. First, it is concerned only with the expected probability of success for <u>borderline</u> examinees, those examinees whose total scores fall near the cutoff score. Item p-values, on the other hand, reflect the performance of all examinees. Second, it is a true measure of the internal consistency of a judge's responses, given that the item response theory model is valid.

The following steps were used to calculate the value of Van der Linden's index for the judges in this study:

(1) IRT item parameters were estimated for a sample of 2500 examinees using BILOG (Mislevy & Bock, 1986). The three-parameter model was chosen because most of the judges had said that, in making their judgments, they had taken into account the fact that some examinees who had not mastered the necessary skills would nevertheless be able to guess the correct answer.

(2) Since examinees with the same raw score often have different ability estimates, raw scores could only be converted to corresponding ability estimates by evaluating the test characteristic function at a large number of points along the ability scale. The test characteristic function is defined as the sum of all the item response functions for the test. Evaluated at a specified value of θ, it gives the expected value of the raw test score for an examinee of that ability level. A short program was written to generate values of the test characteristic function in increments of .05 on the θ scale. Each judge's standard was then linked to the value of θ for which the value of the test characteristic function most closely approximated it. The θ value associated with any given judge's standard will be denoted by θ_c.

(3) For each judge, the item response function for each item was evaluated at θ_c to yield the probability p_i that an examinee of that ability level will give a correct response to the item.

(4) Specification errors were computed as

$$e_i = MPL_i - p_i,$$

where e_i is the specification error of a particular judge on item i and MPL_i is the minimum pass level specified by that judge for item i.

(5) Each judge's specification errors were averaged across items. A given judge's average error will be denoted by E.

(6) The index of intrajudge consistency was computed for each judge using the transformation

$$C \equiv \frac{M-E}{M},$$

where M is the average of the maximum possible errors for the items. Expressed symbolically,

$$M \equiv \frac{\sum e_i}{n},$$

where

$$e_i^{(u)} \equiv \max\{p_i, 1-p_i\}.$$

The transformation makes it possible to compare intrajudge consistency across judges; it indicates the degree to which a judge's average specification error differs from the maximum value it can take on at that judge's value of θ_c. Also, the transformation reverses the scale of measurement so that higher values of C indicate consistency, while lower values indicate inconsistency.

Values of the index were computed for each judge using each of the two standard-setting methods, and results were compared across methods.

Results

The cognitive components method led to a higher standard (35.12 items, or about 64% correct) than that yielded by the Angoff method (30.10 items, or about 55%). Standards for individual judges in Study 2 are presented in Table 2. For 10 of the 12 judges ($.01 < \underline{p} < .05$), the cognitive components standard was higher than the Angoff standard.

Interjudge Variability

The cognitive components method resulted in lower variability at every level, i.e., in the recommended standards, the item MPLs, and the raw responses. As Table 2 indicates, the standard deviations of the judges' standards were 7.50 for the Angoff method and 6.44 for the cognitive components method. The standards resulting from the cognitive components method had a range of about 19 items, as compared to 24 for the Angoff method.

At the item level, too, the cognitive components method resulted in lower variability. Means and standard deviations for item MPLs generated by each method are given in Table 3. As can be seen from the table, the standard deviations of the synthetic MPLs generated by the cognitive components model were lower than those of the corresponding Angoff MPLs for 49 of the 55 items (*p* < .001 using a binomial signs test), and for 2 of the remaining 6 items, the two standard deviations were equal.

Table 2

Recommended Standards by Judge

	Angoff Method	Cognitive Components Method
Judge 1	39.61	43.05
Judge 2	31.11	32.64
Judge 3	29.30	37.42
Judge 4	21.26	27.45
Judge 5	39.30	24.27
Judge 6	33.55	33.07
Judge 7	25.29	42.30
Judge 8	26.66	26.81
Judge 9	15.30	35.50
Judge 10	39.40	43.26
Judge 11	27.75	37.29
Judge 12	32.65	38.33
MEAN	30.10	35.12
SD	7.50	6.44

Table 3

Mean MPLs

	Angoff	Cognitive Components		Angoff	Cognitive Components
Item	Mean (SD)	Mean (SD)	Item	Mean (SD)	Mean (SD)
202	.55 (.23)	.76 (.14)	414	.70 (.23)	.74 (.13)
203	.45 (.23)	.74 (.10)	415	.46 (.21)	.33 (.18)
205	.63 (.21)	.86 (.12)	416	.56 (.29)	.61 (.22)
206	.43 (.20)	.92 (.10)	418	.70 (.26)	.61 (.16)
207	.45 (.19)	.47 (.17)	421	.45 (.18)	.42 (.20)
209	.69 (.15)	.66 (.10)	422	.56 (.21)	.71 (.12)
210	.54 (.20)	.54 (.16)	423	.53 (.23)	.58 (.16)
211	.56 (.23)	.84 (.12)	425	.40 (.24)	.31 (.16)
212	.40 (.26)	.69 (.16)	426	.35 (.19)	.58 (.16)
213	.43 (.20)	.34 (.17)	601	.45 (.22)	.41 (.13)
215	.55 (.26)	.75 (.10)	605	.63 (.23)	.65 (.23)
219	.48 (.19)	.36 (.18)	606	.57 (.28)	.61 (.09)
220	.62 (.23)	.72 (.19)	608	.55 (.19)	.57 (.12)
222	.49 (.22)	.74 (.14)	610	.44 (.29)	.73 (.18)
223	.41 (.23)	.61 (.21)	611	.66 (.22)	.57 (.22)
224	.52 (.26)	.86 (.12)	612	.75 (.25)	.74 (.16)
225	.65 (.15)	.54 (.16)	615	.60 (.30)	.72 (.20)
227	.60 (.22)	.61 (.18)	617	.51 (.24)	.51 (.12)
228	.48 (.20)	.54 (.16)	619	.74 (.31)	.99 (.16)
230	.51 (.21)	.34 (.13)	620	.54 (.29)	.72 (.16)
401	.39 (.23)	.73 (.14)	621	.49 (.23)	.39 (.16)
402	.45 (.21)	.69 (.16)	622	.65 (.27)	.72 (.18)
403	.60 (.24)	.67 (.17)	623	.45 (.23)	.76 (.15)
406	.86 (.10)	.74 (.12)	624	.88 (.12)	.91 (.17)
407	.63 (.22)	.57 (.16)	627	.36 (.24)	.69 (.15)
410	.45 (.25)	.44 (.13)	628	.71 (.27)	.87 (.14)
411	.72 (.30)	.62 (.23)	629	.41 (.22)	.57 (.15)
413	.44 (.27)	.75 (.09)			

Raw responses were also considerably less variable for the cognitive components method, as was the case in Study 1. For Angoff MPLs, item standard deviations ranged from .098 to .309,

BEST COPY AVAILABLE

with a mean standard deviation of .228. The probabilities specified for cognitive components varied less, with standard deviations ranging from .016 to .213 and averaging .122.

Correlations between raw responses for pairs of judges are a further indication of higher interjudge agreement using the cognitive components method. Tables 4 and 5 present the correlation matrices for the cognitive components ratings and Angoff ratings, respectively. Of the 66 correlations between judges, 55 were higher for the cognitive components method than for the Angoff method ($p < .001$). Forty-six of the 66 correlations were significant at the .01 level for the cognitive components data (i.e., the MSRs), while only 25 were significant at the same level for the Angoff MPLs.

Intrajudge Consistency

Intrajudge consistency as measured by the index proposed by Van der Linden (1982) was higher for the cognitive components method than for the Angoff method. Table 6 presents, for each judge, both the average absolute error and the value of the consistency index. Nine of the 12 judges were more consistent using the cognitive components method than using the Angoff method ($p < .10$). The average absolute error for each judge also tended to be smaller for the cognitive components method, ranging

Table 4

Correlation Matrix for MSRs (Cognitive Components Method)

Judge	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	.62	.77	.55	.65	.57	-.08	.56	.60	.60	.53	.61
2		1.00	.63	.52	.64	.41	-.14	.67	.37	.53	.41	.41
3			1.00	.75	.61	.59	-.13	.69	.59	.85	.77	.80
4				1.00	.57	.57	-.11	.46	.47	.67	.70	.64
5					1.00	.49	-.10	.58	.37	.63	.53	.49
6						1.00	-.04	.25	.38	.58	.54	.69
7							1.00	-.01	.32	-.05	-.10	-.13
8								1.00	.52	.59	.45	.48
9									1.00	.57	.47	.53
10										1.00	.69	.76
11											1.00	.79
12												1.00

Table 5

Correlation Matrix for MPLs (Angoff Method)

Judge	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	.55	.64	.50	.42	.38	.15	.28	.23	.34	.36	.25
2		1.00	.56	.43	.24	.21	.18	.37	.03	.30	.13	-.09
3			1.00	.39	.39	.28	.15	.36	.05	.39	.46	.08
4				1.00	.34	.51	.32	.27	-.09	.30	.35	.39
5					1.00	.19	.24	.20	.18	.07	.40	.45
6						1.00	.15	.09	.16	.27	.28	.35
7							1.00	.22	.38	.32	.28	.26
8								1.00	.13	.50	.29	.09
9									1.00	.33	.01	.21
10										1.00	.16	-0.00
11											1.00	.27
12												1.00

BEST COPY AVAILABLE

from .123 to .209, compared with a range of .136 to .262 for the Angoff method. This was true even though the cognitive components method led to a greater average maximum possible error for 10 of the 12 judges. What these results mean is that, for the cognitive components method, judges' MPLs for individual items tended to be more consistent with their "global"

Table 6

Van der Linden's Index of Intrajudge Consistency

	Angoff		Cognitive Components	
Judge	E[a]	C[b]	E[a]	C[b]
1	.1465	.8064	.1234	.8446
2	.2121	.6980	.1674	.7641
3	.2140	.7024	.1487	.7986
4	.2622	.6191	.2093	.7571
5	.2448	.6750	.2051	.7026
6	.1727	.7583	.1805	.7465
7	.2015	.7081	.1309	.8329
8	.1783	.7424	.1823	.7368
9	.1773	.7558	.2031	.7204
10	.1355	.8201	.1282	.8386
11	.1794	.7413	.1817	.7539
12	.2405	.6610	.1539	.7937

[a] Values represent average absolute error of specification.

[b] Values represent consistency as measured by Van der Linden's index.

definitions of a minimally competent examinee as represented by their overall standards.

Profiles of individual judges were examined to explore the relationships among various characteristics of each judge's work, particularly its consistency. For both methods, the two most consistent judges were Judge 1 and Judge 10. Interestingly, for both methods, the standards recommended by these two judges differed from each other by less than .25. Further, the difference between the standards for the two methods was approximately equal for these two judges; the cognitive components standard was higher than the Angoff standard by 3.44 items for Judge 1 and 3.86 items for Judge 10. This result could, of course, be due to chance. Alternative hypotheses, however, are possible. It could be the case, for example, that if a judge is very consistent using both methods, the difference between the two standards approaches some ideal magnitude (in a given direction!) that is determined by the particular cognitive components model used (i.e., the specific set of components).

The two least consistent judges were a different pair of judges for each method, and no particular patterns were found. This is not entirely surprising since inconsistency in general tends to be associated with randomness.

The two judges whose consistency benefited most from using the cognitive components method were Judge 7 and 12, for whom the cognitive components consistency measures were higher than the Angoff values by .13 and .12, respectively. For both of these judges, correlations between MPLs and item p-values were very low (.19 and .12) for the Angoff method, but respectable (.47 for both judges) using the cognitive components method. A curious fact is that Judge 7 was also the judge whose cognitive components responses were correlated least with those of the other judges (*r* = -.04 to -.14). Clearly, the relationships among these types of data are complex, and there is no clear-cut way to assess the "quality" of a judge's work, nor of a standard-setting procedure.

Discussion

The results of this study are extremely encouraging with regard to the potential of the cognitive components model as an alternative approach to standard setting. As in the initial study (McGinty & Neel, 1996), the cognitive components model resulted in lower variability among judges at all levels of the process. This may suggest that a substantial proportion of disagreement among judges using the Angoff method is due to judges' differing abilities to perceive the important characteristics of items. We believe that this type of disagreement is undesirable. In contrast, it is natural, and not necessarily undesirable, for judges to have differing opinions about the level of performance that should be required. By directing judges' attention to the important features of the items, the cognitive components model reduces the former type of variability.

This study also provides some evidence that judges' ratings may be more internally consistent using the cognitive components method than using the Angoff method. Clearly, replication is needed, since the results could have been due to chance. It is encouraging, however, that 9 out of 12 judges were more consistent using the cognitive components method. The probability of obtaining this result by chance alone was about .07, but the power of the significance test was low since there were only 12 judges.

Taken together, the initial study (McGinty & Neel, 1996) and the current one provide rather striking evidence for the potential of the cognitive components approach to standard setting. In addition to its potential advantages in the area of reliability, it may have some practical advantages as well (see McGinty & Neel, 1996). Finally, it offers exciting possibilities for use in combination with other standard-setting methods; these are discussed at some length in McGinty & Neel (1996).

References

Angoff, W. H. (1971). Scales,norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Donders, F. C. (1969). On the speed of mental processes (W. G. Koster, Trans.). Acta Psychologica, 30, 412-431.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93(1), 179-197.

Embretson, S. E. (1984). A general latent trait model for response processes. Psychometrika, 49, 175-186.

Embretson, S. E. (Ed.) (1985a). Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. E. (1985b). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 195-218). New York: Academic Press.

Fischer, G. H. (1973). The linear logistic test model as an instrument in education research. Acta Psychologica, 37, 359-374.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*(4), 461-475.

McGinty, D., & Neel, J. H. (1996). *Judgmental standard setting using a cognitive components model.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp.41-82). Hillsdale, N.J.: Erlbaum.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*, 187-214.

Sternberg, R. J. (1977). *Intelligence, information processing, and analytical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1978). Isolating the components of intelligence. *Intelligence, 2*(2), 117-128.

Sternberg, R. J. (1979). Six authors in search of a character: A play about intelligence tests in the year 2000. *Intelligence, 3*(3), 283-293.

Sternberg, R. J. (1983). Components of human intelligence. *Cognition, 15,* 199-206.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction time experiments. American Scientist, 4, 412-457.

Sternberg, S. (1975). Memory-scanning: New findings and current controversies. Quarterly Journal of Experimental Psychology, 27, 1-32.

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge consistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19(4), 295-308.

Whitely, S. E. (1980). Latent trait models in the study of intelligence. Intelligence, 4, 97-132.

GERA

TM027701

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Interjudge Variability and Intrajudge Consistency Using the Cognitive Components Model for Standard Setting

Author(s): Dixie McGinty, John H. Neel, and Yu-Sheng Hsu

Corporate Source:

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☑ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Dixie McGinty

Printed Name/Position/Title: Dixie McGinty / Assistant Prof.

Organization/Address:
Dept. of Administration, Curriculum, + Instruction
Western Carolina University
Cullowhee, NC 28723-9039

Telephone: 704-227-7415

FAX: 704-227-7388

E-Mail Address: dmcginty@wcu.edu

Date: 8-20-97

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission tc reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2d Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com