DOCUMENT RESUME

ED 371 001 TM 021 621

AUTHOR Chang, Lei; And Others
TITLE Does a Standard Reflect Minimal Competency of Examinees or Judge Competency?
PUB DATE Apr 94
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Economics; *Evaluators; Experience; *Interrater Reliability; *Judges; *Knowledge Level; Minimum Competencies; Minimum Competency Testing; Teacher Certification; Test Construction; *Test Items
IDENTIFIERS Angoff Methods; *Standard Setting

ABSTRACT

The present study examines the influence of judges' item-related knowledge on setting standards for competency tests. Seventeen judges from different professions took a 122-item teacher-certification test in economics while setting competency standards for the test using the Angoff procedure. Judges tended to set higher standards for items they got right and lower standards for items they had trouble with. Interjudge and intrajudge consistency were higher for items all judges got right than items some judges got wrong. Procedures to make judges' test-related knowledge and experience uniform are discussed. (Contains 19 references and 3 tables.) (SLD)

Does a standard reflect minimal competency of examinees or judge competency?

Lei Chang, Charles Dziuban, Michael Hynes
University of Central Florida
Arthur Olson
University of West Florida

Paper presented at the 77th Annual Convention of American Educational Research Association, New Orleans, 1994

Correspondence concerning this article should be sent to Lei Chang, Department of Educational Foundations, University of Central Florida, Orlando, FL 32816-1250.

Abstract

The present study examines the influence of judges' item-related knowledge on setting standards for competency tests. Seventeen judges from different professions took a 122-item teacher certification test in economics while setting competency standards for the test using the Angoff procedure. Judges tended to set higher standards for items they got right and lower standards for items they had trouble with. Interjudge and intrajudge consistency were higher for items all judges got right than items some judges got wrong. Procedures to make uniform judges' test-related knowledge and experience are discussed.

Does a standard reflect minimal competency of examinees or judge competency?

In the past four decades, numerous procedures have been introduced and refined to establish performance standards on criterion-referenced achievement tests (Jaejer, 1989; Cizek, 1993). All of these procedures are judgmental and arbitrary (Jaeger, 1976, 1989; Glass, 1978). They entail, in varying ways, judges' perceptions of how minimally competent examinees would perform on each item of the test. Judgmental errors arise when judges differ in their conceptualizations of minimal competency and, within judges, when such conceptualizations are not stably maintained across items. The motivation behind the four decades of experimenting with different standard setting methods is to reduce these errors or to maximize intrajudge and interjudge consistency in reaching judgements.

What are the possible causes of judgmental inconsistencies both within and across judges? Plake, Melican, and Mills (1991) classified the potential causal factors into three categories in relation to judge backgrounds, items and their contexts, and standard-setting processes. Among the judge-related factors, judges' specialty and professional skills are suspected to influence their item ratings during standard setting (Plake et al., 1991). In many content areas, the domain of knowledge is so broad that it is unrealistic to expect the judges to know everything (Norcini, Shea, & Kanya, 1988) on the test even though they are considered experts. The fact that judges are often

deliberately selected to represent different professional experiences (Jaeger, 1991) makes it more difficult to assume that their domain knowledge in relation to each individual item on a test is a constant but not a variable. Empirical findings of markedly different standards derived by judges of different professions (e.g., Jaeger, Cole, Irwin, & Pratto, 1980, cited from Jaeger, 1989; Roth, 1987) may be explained by the judges' different training and vocational focuses regarding a broadly defined domain of knowledge. Another empirical finding is that judges have different perceptions about minimal competencies (ven de Linden, 1982; Plake et al., 1991). It is logical to suspect that judges' different professional focuses influence their perceptions of minimal competency in relation to an item. To what extent, then, does a competency standard derived for minimally competent examinees reflect the strengths and weaknesses of the judges with respect to the content domain of competency?

To date, only one empirical study has attempted to investigate this question. Norcini et al. (1988) compared three cardiologists with three pulmonologists in their ratings of items representing these two specialty areas. There was no statistically significant difference in ratings between the two groups of three specialty judges. These results, however, are inconclusive for two reasons. First, the independent variable, specialty expertise, was not operationally defined; in other words, there was no objective evaluation of judges' item-related

expertise in each content area. The vagueness of expertise distinction was further muddled by the fact that all six judges were involved in writing and reviewing the items being rated. As the authors admitted, "This experience may have made them "experts" in the narrow domain of the questions on the examination and mitigated the effect of specialization" (p. 60). Other researchers have echoed similar criticism (e.g., Plake et al., 1991).

In the present study, item-related expertise of the judges is operationally defined by having the juduges take the test for which they are to provide competency standard. It is hypothesized that (1) judges will set a higher standard for items they answer correctly than for items they answer incorrectly, and (2) intrajudge and interjudge consistency will both be higher when all of the judges answer all of the items correctly than when some of the judges answer some of the items incorrectly.

Interjudge and Intrajudge Consistency

Interjudge consistency refers to the degree to which standards derived by different judges agree with each other. Intrajudge consistency (ven de Linden, 1982) refers to the degree to which an individual judge's estimate of item difficulty is consistent among items. It is usually evaluated by comparing a judge's estimate of item difficulty with an empirical item difficulty[1], both of which are based on minimally competent examinees. Intrajudge consistency can also be viewed as internal consistency reliability of judge-estimated item difficulties

(Friedman & Ho, 1990). Reflecting Friedman and Ho's definition of intrajudge consistency and the definition of interjudge consistency, Brennan and Lockwood (1980) used generalizability theory to estimate judgment errors both within and across judges associated with the Angoff and Nedelsky procedures. The present study uses Brennan and Lockwood's approach and examines intrajudge and interjudge consistency viewed from the perspective of generalizability theory. The following discusses interjudge and intrajudge consistency within generalizability theory.

X_{ji} indicates a judge's score on a item from the population of judges and universe of items. The expected value of a judge's observed score is $\mu_j \equiv E_i X_{ji}$. The sample estimate is $\bar{X}_j$. The expected value of an item is $\mu_i \equiv E_j X_{ji}$. The corresponding sample estimate is $\bar{X}_i$. The expected value over both judges and items is $\mu \equiv E_j E_i X_{ji}$. The sample estimate is $\bar{X}$ or the cutting score.

X_{ji} can be expressed in terms of the following equation:

$$X_{ji} = \mu + \mu_j\sim + \mu_i\sim + \mu_{ri}\sim$$

where μ is the grand mean,

$\mu_j\sim = \mu_j - \mu$ is the judge effect,

$\mu_i\sim = \mu_i - \mu$ is the item effect,

$\mu_{ji}\sim = X_{ri} - \mu_j - \mu_i - \mu$ is the residual effect.

For each of the three score effects there is an associated variance component. They are:

$$\sigma^2(j) = E_r(\mu_j - \mu)^2$$

$$\sigma^2(i) = E_i(\mu_i - \mu)^2$$

$$\sigma^2(ji) = E_j E_i(X_{ji} - \mu_j - \mu_i + \mu)^2$$

The three variance components are estimated by equating them to their observed mean squares in ANOVA:

$\hat{\sigma}^2(j) = [MS(j) - MS(ji)] / n_i$;

$\hat{\sigma}^2(i) = [MS(i) - MS(ji)] / n_j$;

$\hat{\sigma}^2(ji) = MS(ji)$.

Adding up these estimates of variance components gives the estimate for the expected observed score variance:

$$\hat{\sigma}^2(X_{ji}) = \hat{\sigma}^2(j) + \hat{\sigma}^2(i) + \hat{\sigma}^2(ji) \qquad (1)$$

These variance components are associated with a single judge's score on a single item (X_{ji}). In a standard setting situation, a sample of n'_j judges and n'_i items are used to estimate $\overline{X}$, the cutting score. By the central limit theorem, the variance associated with $\overline{X}$ is:

$$\hat{\sigma}^2(\overline{X}) = \hat{\sigma}^2(j)/n'_j + \hat{\sigma}^2(i)/n'_i + \hat{\sigma}^2(ji)/n'_j n'_i \qquad (2)$$

$\hat{\sigma}^2(\overline{X})$ consists of two components:

$$\hat{\sigma}^2(\overline{X}_j) = \hat{\sigma}^2(i)/n'_i + \hat{\sigma}^2(ji)/n'_j n'_i \qquad (3)$$

$$\hat{\sigma}^2(\overline{X}_i) = \hat{\sigma}^2(j)/n'_j + \hat{\sigma}^2(ji)/n'_j n'_i \qquad (4)$$

Equations (3) and (4) represent intrajudge and interjudge inconsistencies when n'j judges and n'i items are used to estimate the standard, μ. If some items are more difficult than others, the selection of items will influence the judgement for a minimally competent examinee's absolute level of performance. Thus, $\hat{\sigma}^2(i)/n'_i$ is considered intrajudge inconsistency since it has a direct impact on the expected value of a judge, μ_j. $\hat{\sigma}^2(j)/n'_j$ represents interjudge inconsistency because it influences the expected value of an item over judges, μ_i. It

shows that, if judges have different perceptions of minimal competency and/or item difficulties, the selection of judges will change the item difficulty. Finally, $\partial^2(ji)/n'_j n'_i$ contributes both to intrajudge and interjudge inconsistency. Part of the judge-item interaction indicates that differences in leniency or stringency among judges are registered differently on different items. In other words, judges fail to maintain their standards across items. With a single observation for each judge-item combination, the last interpretation is, however, confounded with other unexplainable effects.

Method and Results

The Test, Judges, and Standard-Setting Procedures

The Florida Teacher Certification Examination in Economics was used to examine the influence of judges' item competency. The test contained 122 4-choice items. Seventeen judges were selected from the state to set competency standards for this test. They consisted of certified high school chemistry teachers Education professors, and district supervisors. The teachers had varying years of classroom experience. A modified Angoff (1971) procedure was used. Judges were first instructed about the Angoff procedure. They were then administered the 122-item test. While taking the test, they estimated item difficulty for minimally competent examinees. They were then given their own test scores and Angoff scores, means and frequency distributions of the panel's test scores and Angoff scores, and the mean and frequency distribution of a sample of examinees who took

the test. With these information packets, they engaged in subsequent "Closure with Consensus" discussions. With this technique, the panel was divided into smaller groups to discuss the material and reach consensus on the cut-off score. Having reached consensus within groups, each group sent an emissary to another group to form new groups to continue with the deliberation. This emissary process was repeated until consensus was reached among all judges regarding the passing score. Data reported in this study consisted of the individual judges' initial test scores and Angoff scores before the open group discussion.

G-Study

A random effect j x i crossed design ANOVA was conducted within the whole sample and two subsamples. The whole sample was an Angoff score matrix of 122 items by 17 judges. The two subsamples had Angoff scores from the same 17 judges on a subset of 46 items. In one subsample, the 46 items were ones that all 17 judges answered correctly when taking the test. This subsample will be referred to as the "homogeneous knowledge" sample. The other subsample had a different set of 46 items where each of the 17 judges missed at least 5 items when taking the test. This subsample will be called the "heterogeneous knowledge" sample. Variance components and intrajudge and interjudge inconsistencies were compared among these three samples.

Insert Tables 1 and 2 about here

The G-study results from the three samples are reported in Table 1 and intrajudge and interjudge inconsistencies are reported in Table 2. Variance components, $\hat{\sigma}^2(j)$ and $\hat{\sigma}^2(ji)$, estimated from the heterogeneous knowledge sample were much larger than those from the homogeneous knowledge sample. $\hat{\sigma}^2(i)$ was similar across the two samples. Correspondingly, judgements were more consistent across judges (interjudge consistency) when they knew the answers to all the items on the test. Interjudge consistency was much worse for the items to which judges did not know all the answers. Intrajudge consistency was similar across the two samples although it was still higher for the items judges knew the answers to than those items some of the judges did not know the answers to. These findings supported the hypothesis that lack of content knowledge increases errors in standard-setting.

T-Tests

T-tests were conducted within an individual judge to test the second hypothesis that a judge's standard was higher for items he/she knew the answers to than for items he/she did not know. The t-test compared a judge's average Angoff score, the standard, derived from the items she/he got right when taking the test against the standard based on the items she/he got wrong. For one judge who did not miss any items, such a comparison was not possible. Thus there were 16 t-tests. The results are

reported in Table 3.

Insert Table 3 here

As can be seen from Table 3, for all 16 judges, their Angoff ratings were much higher for items they knew than for items they did not know. Fifteen out of the 16 t-tests were significant, $\alpha<.05$. Apparently, when a judge knew the answer, the judge expected a larger proportion of minimally competent examinees to get the item right than when the judge himself or herself had trouble with the item.

Discussion

The results from this study are straightforward. Judges' domain knowledge related to the items on a test affect standard-setting both in terms of the mean, or the standard, and variance, or errors surrounding the standard. As a matter of common sense, judges tend to set relatively higher standards for items they know and lower standards for items they do not know. The problem is that the standard thus derived reflects not the minimal competency of the examinees as it should, but the competency of the judges.

Judge competency has similar influences on the consistency of the standard. Interjudge inconsistency arises as a result of the heterogeneous competency background of the judges. When some of the judges do not know some of the items, there is more

discrepancy in the standards derived. On the other hand, judgement is more consistent for items to which all judges know the answers.

One implication of this study is that more emphasis should be placed on training judges prior to standard setting. When judges come from different professions and experiences, it is only natural that they have different focuses on the knowledge domain of which the competency test is a sample. Consequently, they may not be uniformly familiar with every item on the test. Item-related training, including having the judges take the test, will make uniform their experience and expertise so as to reduce interjudge and intrajudge inconsistency.

Logically, however, those who initially did not know an item and learned it through training could still be more lenient when judging that item than other items they knew initially. On the other hand, judges who did better on the test initially may be more stringent in rendering standards than those who did worse despite training. Thus, item related training should also be accompanied by specific instructions to guard against setting "judge competency standards" found in this study. Having the judges take the test and providing them with the test information will help in this regard. For example, knowing that 90% of the panel answered the item correctly, a judge who failed the item is likely to change his/her otherwise low estimate of em difficulty which reflecting the judge's lack of item competency. The results of judges' initial tests can also be used to screen

judges by eliminating the outliers.

Findings from the present study also provide clues to the lack of equitability among different standard-setting methods (Andre & Hecht, 1976; Skakun & Kling, 1980; Koffler, 1980; Brennan & Lockwood, 1980; Poggio, Glasnapp, & Eros, 1981; Mills, 1983; Cross, Impara, Frary, & Jaeger, 1984; Jaeger, 1989). Among the different procedures, the Nedelsky method was often found to produce lower standards (Andrew & Hecht, 1976; Shepard, 1980; Skakun & Kling, 1980; Brennan & Lockwood, 1980; Poggio, Glasnapp, & Eros, 1981; Cross, Impara, Frary, & Jaeger, 1984). In light of the present study, the lower Nedelsky standard may be due to the fact that judges' own difficulty with items are more directly tested with the Nedelsky procedure where the judges have to go through all the alternative answers to eliminate the wrong ones. A judge has to evaluate the similarities and differences among the response options (Smith and Smith, 1988) to determine the probability of eliminating the wrong answers. Such a process taxes a judge's knowledge much more frequently than does determining the difficulty of the item as a whole in the Angoff and other procedures. It is likely that a judge who is fairly confident of the answer to the item becomes more doubtful of his/her item-related knowledge when going through each alternative in the Nedelsky method. According to the findings of the present study, the judge's doubt about an item will be reflected in a lower Nedelsky standard.

Quasi-experimental studies can be conducted to further test

the influence of judge's domain knowledge. Specifically, the Nedelsky method can be compared with the Angoff method for judges expected to know the items, e.g., judges who were involved in developing the items, and for judges who are not expected to know all the answers on the test. We anticipate a greatly reduced difference between the Nedelsky and Angoff procedures for the former than the latter group.

It is important to identify the negative impact of judge knowledge on standard-setting. To a certain degree, subjectively derived standards of minimal competency are expected to reflect the competency of the people who derive them. On the other hand, it is unrealistic to expect judges to be uniformly competent with respect to every item on the test. Further research should seek a better understanding of the "judge competency standard" phenomenon and find ways to minimize it.

References

Andrew, B.J., & Hecht, J.T. (1976). A preliminary investigation of two procedures for setting examination standards, Educational and Psychological Measurement, 36, 45-50.

Beuk, C. H. (1984). A compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21(2), 147-152.

Brennan, R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4(2), 219-240.

Cizek, G.J. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30(2), 93-106.

Cross, L.H., Impara, J.C., Frary, R.B., & Jaeger, R.M. (1984). Acomparison of three methods for establishing minimum standards on the national teacher examinations. Journal of Educational Measurement, 21(2), 113-129.

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

Friedman, C. B., & Ho, K. T. (1990). Interjudge consensus and intrajudge consistency: Is it possible to have both in standard setting? Paper presented at the Annual Convention of the National Council on Measurement in Education, Boston.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 485-514). New York: Macmillan.

Jaeger, R.M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10(2), 3-6, 10, 14.

Koffler, S.L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17(3), 167-178.

Mills, C.N. (1983). A comparison of three methods of establishing cuoff scores on criterion-referenced tests. Journal of Educational Measurement, 20(3), 283-292.

Norcini, J.J., Shea, J.A., & Kanya, D.T. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25(1), 57-65.

Plake, B. S., & Melican, G. J. (1989). Effects of item context on intrajudge consistency of expert judgments via the Nedelsky standard setting method. Educational and Psychological Measurement, 49, 45-51.

Plake, B.S., Melican, G.J., & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10(2), 15-16, 22, 25.

Poggio, J.P., Glasnapp, D.R., & Eros, D.S. (1981). An empirical investigation of the Angoff, Ebel and Nedelsky standard setting methods. Paper presented at the Annual Convention of the American Educational Research Association, Los Angeles.

Roth, R. (1987). The differences between teachers and teacher educators when judging the NTE professional knowledge test

to determine a cut-score. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Mobile, Al.

Shepard, L.A. (1980). Technical issues in minimum competency testing. In D.C. Berliner (Ed.), Review of research in education: Vol. 8, (pp. 30-82). Itasca, Ill.: F. E. Peacock Publishers.

Skakun, E.N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17(3), 229-235.

Smith, R.L., & Smith, J.K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 25(4), 259-274.

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

Footnote

[1]The empirical item difficulty can be obtained in three ways. The most straightforward way is to determine the proportion of people getting the item right from a sample of minimally competent examinees (Plake et al., 1991). When such sample is unavailable as often is the case, it can be derived from certain part of the distribution when the test is administered to a total group of examinees (Plake et al., 1991). Finally, it can be mathematically estimated through the application of an IRT model (ven de Linden, 1982; Friedman & Ho, 1990; Plake et al., 1991).

Table 1

Variance estimates from G-studies

Source	df	MS	$\hat{\sigma}^2$
		Total Sample	
Item (*i*)	121	4.480	.24541
Judge (*j*)	16	15.300	.12287
ji	1936	0.306	.30653
		Homogeneous Knowledge Sample	
Item (*i*)	45	2.086	.10675
Judge (*j*)	16	4.329	.08822
ji	720	0.271	.27108
		Heterogeneous Knowledge Sample	
Item (*i*)	45	2.175	.10872
Judge (*j*)	16	6.941	.14376
ji	720	0.327	.32717

Table 2 Judge Competency 20

Intrajudge and Interjudge Inconsistencies

n'_i:	46	80	122	46	80	122	46	80	122
n'_j:	4	4	4	8	8	8	17	17	17
				Homogeneous Knowledge Sample					
$\hat{\sigma}^2(\bar{X})$:	.025848	.024236	.023485	.014084	.012785	.012180	.010524	.006723	.006195
	(7.4)	(7.2)	(7.0)	(5.5)	(5.2)	(5.1)	(4.7)	(3.8)	(3.6)
$\hat{\sigma}^2(\bar{X}_j)$:	.003793	.002181	.000143	.003057	.001757	.001152	.002667	.001533	.001005
	(2.8)	(2.1)	(1.7)	(2.5)	(1.9)	(1.5)	(2.4)	(1.8)	(1.5)
$\hat{\sigma}^2(\bar{X}_i)$:	.023528	.022902	.022610	.011764	.011449	.011305	.005536	.005388	.005320
	(7.0)	(6.9)	(6.9)	(5.0)	(4.9)	(4.9)	(3.4)	(3.4)	(3.3)
				Heterogeneous Knowledge Sample					
$\hat{\sigma}^2(\bar{X})$:	.040082	.038321	.037502	.021223	.019841	.019196	.011238	.010056	.009505
	(9.2)	(9.0)	(8.9)	(6.7)	(6.5)	(6.4)	(4.9)	(4.6)	(4.5)
$\hat{\sigma}^2(\bar{X}_j)$:	.004142	.002382	.001562	.003253	.001870	.001226	.002782	.001599	.001049
	(3.0)	(2.2)	(1.8)	(2.6)	(2.0)	(1.6)	(2.4)	(1.8)	(1.5)
$\hat{\sigma}^2(\bar{X}_i)$:	.037718	.036962	.036611	.018859	.018481	.018305	.008875	.008697	.008614
	(8.9)	(8.8)	(8.8)	(6.3)	(6.3)	(6.2)	(4.3)	(4.3)	(4.3)

Note. Numbers in parentheses are standard errors in terms of number of items.

Table 3

T-Test Results

Judge	Angoff Scores				T-Test
	Items Right		Items Wrong		
	$\bar{X}_j$	n'_i	$\bar{X}_j$	n'_i	
1	.69	86	.53	36	3.08**
2	.58	86	.43	36	2.87**
3	.54	84	.36	38	3.17***
4	.72	107	.60	15	2.42*
5	.79	94	.61	28	5.21***
6	.63	93	.54	29	2.37*
7	.51	86	.40	36	1.97*
8	.71	102	.44	20	5.62***
9	.72	87	.53	35	3.33**
10	.75	91	.59	31	3.62**
11	.61	90	.37	32	5.13***
12	.57	90	.42	32	4.02***
13	.47	68	.24	54	4.41***
14	.40	102	.33	20	1.44
15	.66	107	.48	15	3.87***
16	.70	109	.37	13	5.34***

Note. $\bar{X}_j$ is a judge's mean Angoff rating based on n'_i items for which he/she got right (Items Right) or wrong (Items Wrong) when taking the test.

*$p<.05$, two-tailed. **$p<.01$, two-tailed. ***$p<.001$, two-tailed.