Led

Raimo Konttinen

Using teachers´ knowledge of their students in setting standards

Institute for Educational Research, University of Jyväskylä

Jyväskylä, Finland

Raimo Konttinen

Professor of Research Methodology at the national Institute for Educational Research in Finland

Early 1980's

Using teachers´ knowledge of their students in setting standards

Abstract

The study explores possibilities in deriving cut-off scores (CS) in criterion-referenced testing (CRT) from teacher ratings with Contrasting Groups method. Goals and contents were restricted to those common for the entire age group (core curriculum) in mathematics and mother tongue reading comprehension (Finnish and Swedish) on grades 3 (math 4), 6, and 9. CSs were derived from the logit regression of the teacher ratings on test score and on some explanatory variables. CSs always covaried significantly with some explanatory variable(s), most notably with sex (CS lower for girls) and mean test score of class (CS highest in the best classes), but only occasionally with other variables (class size, class type, urbanization of the school environment, study program, and bilingualism). Despite the fact that invariant CSs could not be obtained, the Contrasting Groups method and the logit regression can be useful analytic tools, and uses of the procedure in evaluation ~~is~~ are discussed.

Prepared probably for IEA or perhaps Scandinavian Journal of Educational Research; no information if the manuscript has been published

Criterion referenced testing (CRT) is used in the educational context to find out the content and amount of a student´s attainments or as Popham (1978, 93) defines it: "A criterion-referenced test is used to ascertain an individual´s status with respect to a well-defined behavioral domain." Although the specification of a domain is not always an easy task, the description of the status may be even more problematic. When the score is used in decision making, some indication of the sufficiency of the attainments or a standard may be needed. It can be expressed in the form of one, possibly several cut-off scores, and several ~~numerous~~ methods for deriving them have been developed (Glass, 1978; Hambleton, Swaminathan, Algina & Coulson, 1978; Hambleton, 1980; Shepard, 1980a). There is, however, much controversy on both the validity of the cut-off scores and ~~of~~ their usefulness. The Present study investigates the variability of cut-off scores in one of them, the Contrasting Groups method developed by Zieky and Livingston (1977).

In her review of standard-setting methods, Shepard (1980b) makes a distinction between student diagnosis, student certification, and program evaluation. The Results of the present study are pertinent to the last mentioned use of cut-off scores. The study is related to ~~It was conducted in connection of~~ an evaluation of national core curricula for mathematics and reading comprehension for grades 1 through ~~to~~ 9 where also an assessment of the minimum acceptable achievement in each area tested was studied ~~desired~~. Standard setting has mainly been studied in connection with ~~of~~ individual level decision making and in program evaluation the problems and usefulness of cut-off scores have ~~has~~ not been much investigated.

Validity of cut-off scores and the evaluation context x)

The validity of the results concerning a standard setting method depends on, whether the derivation and use of the cut-off scores can be justified by the educational evaluation and decision making context. ~~in question~~. In analyzing the context, it was seen useful to conceive of ~~the~~ standard setting as a process consisting of three stages: judg~~e~~ments, collection of the judg~~e~~ments and decision on the standard, and the use of the standard in solving some educational decision making problem. Each stage has different problems: (1) How to collect the judg~~e~~ments? (2) How to derive the standards to make them useful in solving the educational problem in question? (3) What is the appropriate use of cut-off scores in solving the problem?

The present empirical study is related to the second problem; the possibilities of ~~to~~ summariz~~e~~ing experts' opinion by a cut-off score. The starting point has been, that the results would not have ecological validity, if the first and third problems were ~~are~~ left unanswered. Any study of standard setting methods should, from this point of view, be conducted in a~~n~~

context where the standards are used in solving a realistic educational problem and the task for the judges is ~~connected~~ related to this problem. It is not obvious, however, that the two requirements can be met in ~~connection of~~ program evaluation.

As to ~~the~~ problem (3) above, Shepard (1980b) does not see much use for dichotomous standards in program evaluation: "Because standards impose an artificial dichotomy, they obscure performance information about individuals along the full performance continuum" (p. 464). On the other hand, Livingston (1980) points out that knowing the strengths and weaknesses of the students at the cut-off point can be illuminating. ~~This~~ the latter view reflects also the thinking of the present study. If the curriculum is to be taken as a goal for an institution adopting it, it is of interest to know~~,~~ what the least acceptable achievements are. If they are judged as insufficient or if the number of students below the standard is thought to be too great, Some measures to change the situation would probably be considered ~~otherwise good results would not be a sufficient~~ argument for not irrespective of whether the results were even worse in other student groups. ~~taking any measures to~~ change the situation.

It is not intended to maintain that standard setting always is useful in program evaluation. There are certainly situations, where the educational problem can very well be solved without cut-off scores (see Glass, 1978, for some examples). However, judgement of the usefulness of cut-off scores must be based on a thorough analysis of each specific educational problem in question. In analyzing a specific case, evaluation models, like

Stufflebeam's (1971) CIPP-model, might offer fruitful frameworks, but, with the exception of ? ~~in addition to~~ the Shepard's (1980b) work, they have not been used much in connection of standard setting debate.

The task and instructions given to the judges (problem (1) above) also determines the validity of the obtained cut-off scores. Shepard (1981b) illustrates the problem in ~~connection of~~ the Nedelsky~~'s~~ method. Also, the validity of the cut-off scores obtained by the Contrasting Groups method would suffer if, say, the teachers are asked to name the students who have the attainments that represent the ideal aspirations of the school system and the test is used to sort out students who have attained reasonable minima only. The Instructions to the judges determine whether a judge can use the standard setting procedure in a technically correct way exact. From the above example already it is evident that more important from the point of view of the validity of results is, however, what is the conception of the educational problem that the instructions convey.

Livingston (1980) lists four characteristics of a good standard setting method and all of them are related to the compatibility of the judges' task with the use of the results: (1) Judg~~e~~ments must be made in a way that is meaningful to the persons who are making them, (2) the process must take into account the purpose for which the test is being used, (3) judg~~e~~ments must be made by persons who are qualified to make them, and (4) the process must take into account the consequences of both types of decision errors. In the present study it seemed possible to formulate the judges' task ~~in congruence~~ so that it was congruent with the purpose of the test use. Teachers were asked to rate ~~of~~ each of ~~her~~ their students, in terms of whether he or she had attained a~~n~~ level, ~~on~~ in the domain represented by the test, that in her opinion could be set as the common goal for the entire age group. Ratings were asked for the purposes of curriculum evaluation and these were not used in any decisions concerning individual students. In giving their ratings the teachers were in fact addressing to the same question as the researchers of the evaluation project, only in a different language, through concrete examples taken from the classroom (individual students whom they knew well).

The relation of standard setting to evaluation and decision making problem has been emphasized. It can be argued that clarification of this role is a prerequisite, but not a guarantee, for valid results. Even though part of the variability in the cut-off scores can be reduced by a proper design of a standard setting experiment, there remains many other potential sources of error. These ~~are~~ constitute the problem of the empirical part of the study.

Variability of the cut-off scores

Previous research has shown variation both between standard setting methods as well as within methods (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Koffler, 1980; Skakun & Kling, 1980;

Saunders, Ryan & Huynh, 1981) Variation has been studied using the methods of Angoff (1971), Ebel (1972), and Nedelsky (1954), but in the light of the evidence on them, there is ~~all~~ reason to assume variation also in the cut-off scores produced by the Contrasting Groups method.

The variation in the cut-off scores can be assumed to be attributable largely to some systematic effects rather than to random errors, but there is only little research on~~t~~ what they might be. One exception is the study by Brennan and Lockwood (1980), in which they analyz~~ing~~ed the variance components of the ratings of 126 items from five raters, each using both the Nedelsky and Angoff methods. The between-procedures variance component was substantial as could be expected on the basis of other studies, and it was greater than the between-raters component. The differences between raters were, however, also great and several times the residual variance component which included rater x item interaction and errors. Expressed in the form of reliability coefficients, the cut-off score of a rater from 126 items could be determined with a high reliability, .85 in the Angoff method and .93 in the Nedelsky method.

Koffler (1980), using the Contrasting Groups method, found in one case, 11th grade Mathematics, so much variation in the test scores of students rated masters and non-masters by the teacher, that the two groups could not be separated at all. The result indicates that there may be as much systematic difference between teachers in judging their students´ mastery as between persons rating items. However, ~~in addition~~ besides to the Brennan and Lockwood study, little is known of the factors that could explain the differences in teachers´ standards.

In the present study, three types of sources of variation in the cut-off scores were assumed: context and process of decision making, and achievement measurements (Figure 1).

======================

FIGURE 1 SOMEWHERE HERE

======================

In the Contrasting Groups method, like in any other method, which derives the cut-off scores from the relationship between ratings of the students and their achievement test scores, the measuring characteristics of the test constitute one source of variation. In the present study these effects could not, however, be estimated, but the production of the criterion-referenced tests was designed to minimize them.

The context of the decision making in Figure 1 includes (1) the written curriculum and (2) teacher´s conception of it, (3) the domain of observations on which a teacher bases her ratings, and (4) similar domain of the test items. Pe Validity of the cut-off scores depends of the degree on overlap between these four (~~ovals~~ circles in Figure 1). The overlap could not be studied empirically in this study, but an attempt was made to take it into account in planning the study. Test items were designed to cover all the main ~~parts~~ sections of the written curriculum and take into account the contents of teaching materials ~~were~~, which were based on one and the same written curriculum; teachers were familiar with the core curriculum, and they also knew the test items. Still, the fact remains that the context of decision making may produce variation in the cut-off scores.

The empirical study was restricted to two factors related to the process of teachers´ decision making: biasing factors and frame-of-reference factors. These can be identified with within teachers and between-teachers differences, respectively. Bias refers to a teacher´s use of different standards with different students or student groups. Frame-of-reference affects all the ratings of a teacher. In this study, student´s sex and bilingualism were considered as potential biasing factors, whereas the urbanization of the school´s surrounding, the average achievement level of the class, class size, student composition of the class (normal or mixed age group class), and the level of study program (level of courses) were thought to be the most important frame-of-reference factors. The main problem of the study was to find out how much ~~do~~ the cut-off scores vary depending on these explanatory factors.

METHOD

Curricula and school subjects

From the national frame curriculum for the comprehensive schools (grades 1 to 9, age groups 7 to 16), a core curriculum was derived at the National Board of General Education on the basis of teachers´ experiences. The researchers who developed the achievement tests for the monitoring study participated in this definition work.

The goals and contents of the core curricula were designed for the needs of the majority of the age group. They were intended to give all students a good common basis for studies after the ninth grade. To find out the suitability of the core curriculum proposals, a national monitoring study was carried out in the spring of 1979 in mathematics, mother tongue (Finnish and Swedish), and in foreign languages (English and Swedish). The present study is based on the results on the mathematics tests and of the reading comprehension tests in Finnish and Swedish as mother tongue.

Subjects

National samples stratified by ~~the~~ degree of urbanization (towns versus others) and by the size of school were drawn separately for each subject area studied. The study was carried out on three grade levels: third (grade four in mathematics), sixth and ninth.

As the tests were given at the end of the school-year, the average age of students was about 10 (mathematics 11), 13 and 16 years respectively in the three samples.

Achievement tests

For mathematics and reading comprehension, pools of 150 - 200 items were derived based on the goals and contents stated in the national core curriculum proposals. Items were divided and given in booklets (forms) of 30 - 40 partially overlapping items. The tests were as follows.

Mathematics (Math). Each booklet contained 30 completion and multiple-choice items. Alpha reliability coefficients varied between .83 and .88.

Reading Comprehension - Finnish (RC-F). Each test form consisted of two paragraphs chosen randomly from a defined set of Finnish publications (from years 1973 - 76), both followed by 10 to 22 multiple choice questions. The alpha reliabilities of the forms varied from .59 to .92.

Reading Comprehension - Swedish (RC-S). As above, but the texts were deliberately chosen to cover the text types mentioned in the core curriculum proposal. Alpha reliability coefficients for test forms were between .74 and .89.

The scores of the test forms were equated by the one-parameter logistic model with the aid of LOGIST (Wood, Wingersky & Lord, 1976). The fit of the model was studied by a program resembling in this respect Wright´s CALFIT (Wright &

Mead, 1975). Ordinary item analyses using the latent trait values as criterion were also carried out. None of the items were found to conflict with the domain specification or good practices to the extent that it had been red-flagged. The test scores in the subsequent analyses are the latent trait estimates obtained from the 1-parameter LOGIST runs with mean and standard deviations [illegible] approximately 0 and 1, respectively.

Teacher ratings

Teachers were asked to rate each student; whether she or he had attained the common goals and contents that in the teacher's opinion should be required from all or practically all students of this age and grade level (3, 4, 6 or 9). It was pointed out that one formulation of ~~the~~ such a core curriculum could be found in the National Board of General Education's proposal, but it was also emphasized that the teacher should feel free to apply what she or he saw as the common curriculum and performance standards.

In mathematics the teachers were asked to consider the whole subject matter area, but in mother tongue the ratings were asked only considering the reading comprehension skills. In all cases the rating scale was dichotomic: the student had not attained the common goals (0), or the student had attained the common goals (1). Teachers were allowed to ~~give~~ indicate with a question mark ~~for for~~ any student she did not know yet. These cases, about 5 percent of students, were ~~disregarded~~ eliminated from analyses. However, in Swedish speaking area and in the RC-S teachers were asked to give question mark whenever s/he was not sure of her rating. These ratings were accepted in analyses as non-mastery ratings (zeros).

Explanatory variables

Urbanization. Differentiates towns from the less urbanized areas.

Class mean. Average test score of students in the class.

Class size. Number of students in the class. In mixed age groups classes only students from the grade level in question were counted.

Class type. Relevant on grades 3, 4 and 6 only, where there are both normal classes and classes where students from two or more grade levels are taught together (mixed age groups classes).

Sex of student.

Bilingualism. ~~Separates~~ distinguishes students whose both parents have Swedish as their mother tongue from others. The dichotomy was used only in connection of the RC-S, as there are also bilingual students in the schools for Swedish speaking children.

Study program. Relevant in 9th grade mathematics only. There are three programs (sets) of varying coverage and depth on the 8th and 9th grades and two on the 7th grade. It is not very common to change from one set to another. The differences in achievements between the sets are considerable. The Average proportion correct scores were .68, .45, and .26 for sets A, B, and C, respectively.

Statistical analyses

Logistic model (Haberman, 1978) or more specifically multiple binary logistic regression analysis (Anderson, 1980) was used to obtain the cut-off scores and to discover the effects of the explanatory variables on them. Without the effects of the explanatory independent variables the model can be expressed as follows.

$$\text{logit} = \log\left(\frac{P}{1-P}\right) = b_0 + b_T T, \qquad (1)$$

where logit is the natural logarithm of the student's odds and P his probability of getting a favorable (mastery) rating, ~~b0~~ is the intercept or Grand Mean, and bT is slope parameter. T is for the achievement test score.

b_0

b_T

The most interesting point on the test score continuum is the value at which the probability of a student getting a favorable rating is 0.5, i.e. the threshold were the probability of the positive (mastery) rating becomes greater than the probability of the negative (non-mastery) rating. When P is set equal to 0.5 in equation (1) above, logit is zero and solving for T gives T = -b0/bT. This can be taken as the cut-off score based on the teacher ratings.

$-b_0/b_T$

The slope parameter shows whether the ratings can be described in terms of the test score. If the parameter is zero, there is complete uncertainty of the cut-off score. The bigger the slope parameter is, the sharper the distinction between the masters and non-masters becomes. The slope parameter can also be related to the biserial correlation between the ratings and the test score (Lord & Novick, 1968).

The cut-off scores may vary among the teachers and students according to some discrete variables, such as sex, or according to some continuous variables, like the average test score of the class. The generality of the cut-off score can be studied with the multiple logit model by adding to the right hand side of the model (1) those main effects and interactions of the explanatory variables which increase significantly the fit of the model:

$$\text{logit} = b_0 + b_T T + \sum b_j X_j \qquad (2)$$

X_j An explanatory variable, Xj can be either a single (main effect) variable (continuous variable or a dummy variable representing one value of a discrete variable), or a product of several b_j variables (interaction). bj is a parameter and describes the X_j effect of variable Xj. In this case the cut-off score (CS) has to be estimated separately for each value combination of the explanatory variables as follows:

$$CS = \frac{-(b_0 + \sum b_j X_j)}{b_T} \qquad (3)$$

The Cut-off score determined from the logistic regression equation is independent of the distribution of test scores and of the explanatory variables, provided that cases are not selected on the basis of the teacher ratings. Ratings could also be asked after the testings and only on a sample of the students representing two or a few scores. On the other hand, the obtained cut-off scores do not minimize the number of false masters and false non-masters like the use of a discriminant function in Koffler´s (1980) study.

Logistic regression equations were fitted with the computer program GLIM (Barker & Nelder, 1978) with logistic link function and binomial errors. A series of analyses were carried out ~~from~~ on the ungrouped data to find out the independent variables, main effects and two-factor interactions, which improved the fit of the model ($p < .01$). The Chi square test of fit obtained from ungrouped data with one observation in cell is not reliable. The differences between consecutive hierarchical models can, however, be tested with the change in the Chi square and associated degrees of freedom (Haberman, 1978). The Sufficiency of the obtained final models was checked with additional analyses from grouped data. Continuous variables of the final models were recoded into fewer (test score in 4 or 11, and class mean in 3) value classes to avoid small cell frequencies, and the Chi square test of the fit of the final models ~~were~~ was estimated from their crosstabulations.

RESULTS

Predictability of teacher ratings

Final models from the analysis of the ungrouped data are given in Table 1. It includes all the main effects and interactions that improved the fit at $p < .01$ level. When the final models were checked with grouped data, none of the Chi squares was statistically significant at $p < .01$ level. The final logistic regression models are, then, sufficient to explain the between cells variation in the proportion of mastery ratings in the grouped data.

======================

TABLE 1 SOMEWHERE HERE

======================

The fit can be illuminated by comparing teacher ratings with their predicted values. Fitted values, i.e. log odds for getting a positive teacher rating, were calculated from the final models. According to ~~the~~ equation (2), e.g. in the Reading Comprehension - Finnish, 3rd grade, the log odds from Table 1 are

logit(boys) = .76 + 1.10*(Score) - 1.03*(Class mean)
+ 0 + 0*(Score), and

logit(girls) = .76 + 1.10*(Score) - 1.03*(Class mean)

+ 1.46 + .79*(Score).

From the logits, the student's probability for a positive teacher rating was calculated and rounded (one for mastery and zero for non-mastery). The degree of agreement between the actual teacher ratings and the dichotomic predicted rating is given in Table 2.

======================

TABLE 2 SOMEWHERE HERE

======================

It varies from 75 to 88 percent and seems to be of about the same size as in ~~the~~ Koffler's (1980) study.

The estimation of the cut-off score is based on the assumed relationship between teacher ratings and test score. The association between these is strong in all cases. Test score is included in all final models and its parameter is always several times the standard error. This could already be seen from the biserial correlation coefficients of the teacher ratings ~~for the Reading~~ with the test score. Their average was .67 (range .50 - .80), which is comparable to a typical, approximately average item-test biserial correlation of achievement tests.

======================

FIGURE 2 SOMEWHERE HERE

======================

Typical values of the slope parameter are illustrated in Figure 2. It displays the regression of the probability of a positive teacher rating on the test score assuming average class mean (zero), for the Reading Comprehension - Finnish, 3rd grade. As the final model includes sex, separate curves were drawn for boys and girls. However, for the class mean, which is also included in ~~th~~ the final model, only the average value (zero) was given. It is obvious that the location of the CS in not exact. If it would be changed a few tenths of standard deviation, the probability would still be around .5. With fallible test scores this may easily happen ~~is possible~~ and the low reliability of some forms of the RC - F, 9th grade, may explain some of the differences between the results from RC - F and RC - S. However, with tests of good reliability, the slope parameter, i.e. the closeness of the relationship between teacher ratings and test score, does not seem to be a serious source of error in the cut-off score. More problematic is that there usually are several relatively clearly defined cut-off scores, which is also illustrated in Figure 2 and will be discussed next.

Factors affecting the cut-off scores

heading!

In all nine analyses, the cut-off point covaried with at least one explanatory variable, mostly with Class mean and Sex. The effect of the Class mean was found in every analysis of the lower grades (3rd, 4th, and 6th), but in none of the ninth grade analyses. This might be related to the smaller class size at the lower grades, where half of the classes may consist of less than ten students from the grade in question. The possibility was checked by performing the analyses with classes of ten or more students only. The results were, however, more or less the same. The smallest classes seemed only to weaken some of the obtained effects, ~~if anything~~ or none at all. Class mean always had a negative effect, meaning that a teacher rating of a student is not only affected by the absolute value of the test score, but also by its relative standing in the classroom, this again is better in a class with lower class mean.

Most of the teachers had, over the years, taught several classes and they also knew well the curriculum, but still their standards varied with the level of the present class. Results from the 9th grade could be interpreted as indicating that recent experience in several classes may diminish the frame-of-reference effect. At the ninth grade, practically all the teachers of mother tongue and mathematics teach several classes and probably have better knowledge of the degree of variation in students' attainments than teachers in lower grades ~~level teachers.~~ It may also be that teachers were not mislead by the level of the class. They simply gave their best estimate of what was asked, of the level that could be set as the goal for an entire age group. Teachers may have been convinced that given similar students and same opportunities, higher standard would not be reasonable. This interpretation would direct interest to the factors producing

[illegible]

different results, not ~~the~~ differing standards.

The Sex of the student had a significant effect in reading comprehension almost in ~~every~~ all (in 5 of 6) cases at all grade levels. The Cut-off score was always lower for the girls than for boys. If a boy and a girl have an equally weak test score, it is likely that only the girl receives a positive teacher rating. The teachers were not misle~~a~~d by the generally better achievements of the girls in the same way as by the class mean. Sex as a biasing factor seems to be functionally different from the frame-of-reference effect of the class mean. However, rating bias is not the only possible interpretation of these differences. It is also possible that teachers´ threshold for rewarding is lower for girls than for boys in other situations, too.

In mathematics, the sex of student was only weakly related to the cut-off score. Sex differences in the cut-off scores may depend on to what extent teachers have the opportunity to observe the behavior on which the ratings are based. Mathematics teachers receive much evidence on exactly the kind of tasks that the test consisted of. Teachers of mother tongue, on the other hand, do not use comprehension test items routinely.

Other effects were found in some of the analyses only. In the Swedish speaking area, the teachers of ninth grades ~~from~~ in rural areas set more demanding reading comprehension standards that their colleagues in towns. To attain the teachers´ standard in mathematics is more difficult for a sixth grade student, if also students from other grades are taught in the same class (mixed class) than when it consists of sixth graders only (normal class).

Ninth grade mathematics is the only case where students could choose between three more and less demanding study programs and it seems to affect teachers assessment of students. Only, the sign of the effect is seemingly in contradiction with the class mean effects. It could have been expected that teachers of the A-set have highest standards like the teachers of the best classes in general. On the contrary, the teachers of the most demanding study programs had lowest standards. Perhaps these teachers, in attempting to consider realistically the entire age group, assumed the achievements and potentialities of the lower set students to be weaker than they were, or conversely, overestimated the excellence of their own students. This, again, indicates the difficulties of taking into account some other frame-of-reference than that offered by the teacher´s own class.

The effects of the explanatory variables were mainly direct. Only two interaction effects were found. In Reading Comprehension - Swedish, sixth grade, the effect of the Class mean was negative, but only in normal classes (-2.03). In mixed classes it was near zero (-2.03 - 1.91 = -.12).

Reading Comprehension - Finnish, third grade, represents the only example of a case where the steepness of the logit regression varies with some explanatory variable: teachers seem to be better able to make the mastery ~~/~~ or non-mastery decision on ~~the~~ girls than on ~~the~~ boys, and their ratings are more closely related to the measured reading achievement among girls (slope 1.89) than among boys (slope 1.10). It may be that girls produce more evidence of their comprehension of texts than boys or that some other factors than reading comprehension affect teacher´s assessment of boys.

Variation and use of the cut-off score

The Variation of the cut-off scores can be illuminated by calculating them for different value combinations of the explanatory variables in the final models. The effect of the class mean can be considerable, if the sample consists of exceptional classes. When the score range containing about 90 percent of classes was considered, the effect of the class mean on the cut-off score was less than plus/minus .5 test score standard deviation. In the Reading Comprehension - Swedish, sixth grade, the class mean alone produced a standard deviation of about 9 in the cut-off scores, which was biggest class mean effect. The differences between girls´ and boys´ cut-off scores varied from zero to 1.26. The latter is for Reading Comprehension - Finnish, ninth grade, where girls´ cut-off score is -(1.39 + 1.11)/.88 = -2.84 and that for boys -1.39/.88 = -1.58.

Other factors in the final models also created variation in the cut-off scores. Depending on the choice of judges, classes and pupils, widely differing cut-off scores can be obtained. At least part of the effect of the two biggest sources of variation, sex and class mean can be controlled by ensuring that the raters know the domain of curriculum to be rated, have experiences of the pupil behavior indicative of the attainment of the objectives, and that they have had opportunity to come to know all the variation in pupils´ achievement. It would still be safer and more useful to include the most likely biasing and frame-of-reference factors into the data collection design and find out their actual effects.

In program evaluation the cut-off scores become interesting only after seeing their implications. The Number of students above the cut-off score can be simply calculated from the score distribution. If item characteristic curves are known, it is also easy to estimate item difficulties and any sub-domain average (Lord, 1980). The same method can be used to inspect average test outcomes for any score group. However, item and sub-domain means at the cut-off score have special interest among them. They reveal attainments excluding clear failures as well as excellent achievements and focus on those cases where the education, student´s work and the results were acceptable ~~considering~~ taking into account the availability of time and resources in the eyes of the teachers. There may be areas of the curriculum, where the attainments of an average student are tolerable, but the quite acceptable group around the cut-off score likely to remain unnoticed may have too many learning difficulties to benefit ~~of~~ from the subsequent instruction.

Item characteristic curves can be used to obtain item and sub-domain means also for the various cut-off scores in the case where they differ according to some explanatory variable like in the present study. If an average cut-off score is needed, it can be calculated as follows. Continuous variables in the final model (like Class mean) is given its average value. Cut-off scores are then calculated for all value combinations of the discrete factors present in the final model and averaged. E.g. for Reading Comprehension - Finnish, 3rd grade, the cut-off score in an average class (Class mean = 0) is for boys -.69 and for girls -1.17 giving the average -.93 in Table 2.

For illustration, the number of students above the cut-off score and the estimated average item difficulty (proportion correct) are calculated for each test in Table 2. In reporting the evaluation study, these and other similar results were seen as good starting point for a closer critical analysis of the school teaching, rather than as facts. For this purpose, reporting a result was complemented with alternative interpretations. When the outcome was good, it was proposed that teachers may have a too low standard or they do not know well enough how to gauge the attainments in question, or the result is even too good and a result of an undue devotion of time and resources on it. A low ~~prestation~~ performance, on the other hand, might indicate unrealistic aspirations among teachers, teachers´ ignorance of students´ factual attainments, or that the specific subject matter area in question should be treated already at lower grades to familiarize students to it or move ~~at~~ it to later grades because students do not have yet enough readines to study it.

CONCLUSIONS

Teachers as judges in the Contrasting Groups method developed by Livingston and Zieky may have greatly varying standards in mind in giving their ratings. This supports the results of earlier studies. However, the between-judges variation ~~of~~ in the cut-off scores was not random. The differences in the cut-off scores could be explained by a few factors, which either brought bias into the ratings, like sex of the student, or determined the entire frame-of-reference of a teacher, like class mean.

In the light of the present study it seem unlikely that the Contrasting Groups method could produce unambiguous cut-off scores. The effects of biasing factors and frame-of-reference factors could, perhaps, be decreased by proper instructions ~~and~~, training and selection of judges, but probably some judgement is needed to reconcile the deviations in the standards.

Even though the Contrasting Groups method cannot offer ~~one~~ a single objective standard for a test, it can aid and illuminate evaluation and decision making in several ways. (1) If an educational problem can be formulated in a form suitable for the Contrasting Groups method, the procedures described in this report can be used to summarize judges' views and to describe the factors related to them. (2) If some factors affect the cut-off scores, they may as such reveal important aspects of the judgement process. (3) Factors affecting the cut-off scores can also help in reformulating the standard setting problem and raise issues to be decided on before fixing the standard at all. (4) The cut-off scores are, perhaps, most useful in a longer process of evaluation and decision making. After analyzing the evaluation task, the Contrasting Groups method may turn out to be an excellent way to summarize experts' opinions. Logistic regression can then be used to evaluate the ~~unanimosity~~ agreement among the judges and the factors affecting their standards. Also, the meaning of the various cut-off scores may be ~~given~~ illustrated by describing the number of students above the cut-off score and the average achievements of students at the cut-off score. The Reporting of the results with an interpretation of the implications could be useful practice in decision making as well as in a wider discussion.

The complexity and great variety of the evaluation and decision making situations where standards may be needed, and the dependence of the cut-off scores on biasing and frame-of-reference factors warrant the conclusion that some kind of research component should be added to any major new standard setting task. The details of the results of this report are at least partly specific to the educational and evaluation setting of the present study. However, the type of analysis carried out here might offer one possible model of a useful research component that would be relatively easy to implement.

REFERENCES

ANDERSON, E.B. Discrete statistical models with social science applications. Amsderdam: North-Holland Publishing Co., 1980.

ANDREW, B.J. & HECHT, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 45, 4-9.

ANGOFF, W.H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational measurement. Washington, DC.: American Council on Education, 1971.

BARKER, R.J. & Nelder, J.A. The GLIM system, release 3. Generalized Linear Interactive Modelling. Manual. Harpenden: Rothamsted Experimental Station, 1978.

BRENNAN, R.L. & LOCKWOOD, R.E. A comparison of the Nedelsky and

Angoff cutting score procedures using generalizability theory. Journal of Educational Measurement, 1980, 17, 167-178.

EBEL, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice Hall, 1972.

GLASS, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

HABERMAN, S.J. Analysis of qualitative data. Volume 1: Introductory topics. New York: Academic Press, 1978

HAMBLETON, R.K., SWAMINATHAN, H., ALGINA, J & COULSON, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

HAMBLETON, R.K. Test score validity and standard-setting methods. In R. A. Berk (Ed.) Criterion-referenced measurement: The state of art. Baltimore, MD.: The Johns Hopkins University Press, 1980.

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

LORD, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ.: Earlbaum.

LORD, F.M. & NOVICK, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

LIVINGSTON, S.A. Comments on criterion-referenced testing. Applied Psychological Measurement, 1980, 4, 575-000.

NEDELSKY, L Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

POPHAM, W.J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-hall.

SAUNDERS, J.C., RYAN, J.P. & HUYNH HUYNH A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 1981, 5, 209-217.

SHEPARD, L. Technical issues in minimum competency testing. In D. C. Berliner (Ed), Review of research in education (Vol. 8). Itasca. IL: F.E. Peacock Publishers, 1980(a).

SHEPARD, L. Standard setting issues and methods. Applied Psychological Measurement, 1980(b), 4, 447-467.

SKAKUN, E.N. & KLING, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

STUFFLEBEAM, D.L. et al. Educational evaluation and decision making. Itasca, Ill.: Peacock, 1971.

WOOD, R.L., WINGERSKY, M.S. & LORD, F.M. LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N.J.: Educational Testing Service, 1976.

WRIGHT, B.D. & MEAD, R.J. CALFIT: Sample-free item calibration with a Rasch measurement model. Statistical Laboratory, Department of Education, The University of Chicago, Research

Memorandum No 18, 1975.

ZIEKY, M.J. & LIVINGSTON, S.A. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, N.J.: Educational Testing Service, 1977.

AUTHOR

KONTTINEN, RAIMO. Address: Institute for Educational Research, University of Jyväskyä, Seminaarinkatu 15, SF-40100 Jyväskyä 10, Finland.

Title: Professor.

Degrees: BA, MA, Ph.D. University of Jyväskylä.

Specialization: Educational measurement, research methodology.

Table 1.
Final models

Effect	b	s(b)	b	s(b)	b	s(b)
READING COMPREHENSION - FINNISH						
	Grade 3		Grade 6		Grade 9	
Grand mean	.76	.12	.86	.11	1.39	.13
Score	1.10	.14	1.50	.12	.88	.10
Class mean	-1.03	.27	-.96	.23	0	
Sex(girl)	1.46	.23	1.33	.18	1.11	.21
Score x Sex(girl)	.79	.26	0		0	
READING COMPREHENSION - SWEDISH						
	Grade 3		Grade 6		Grade 9	
Grand mean	1.05	.10	1.09	.17	2.40	.25
Score	1.48	.13	1.39	.13	1.97	.18
Class mean	-.91	.25	-2.03	.52	0	
Sex(girl)	0		.70	.20	1.43	.26
Classtype(mixed)	0		-.01	.20	-	
Urbanization(rural)	0		0		-1.31	.26
Clmean x Cltype(mixed)	0		1.91	.60	0	
MATHEMATICS						
	Grade 4		Grade 6		Grade 9	
Grand mean	2.85	.22	3.46	.29	1.21	.19
Score	3.11	.27	3.58	.29	2.29	.15
Class mean	-.13	.03	-2.33	.38	0	
Sex(girl)	0		0		.33	.15
Classtype(mixed)	0		-1.08	.26	-	
Study program(B)	-		-		.24	.18
Study program(A)	-		-		.66	.27

Note.
b is estimated effect and s(b) its standard error.
0 is fixed zero, i.e. the effect is not included in the model. indicates
- ~~in place of an estimated effect means~~ indicates, that the explanatory variable is not relevant in this case.

Table 2.

Cut-off scores based on the final models and some desciptive based on them

	RC - Finnish			RC - Swedish			Mathematics		
	3	6	9	3	6	9	4	6	9
Estimated average cut-off score (ACS)[a]	-.93	-1.02	-2.21	-.74	-1.05	-1.21	-.92	-.82	-.73
Average item difficulty at the ACS	49	59	40	50	58	44	48	45	30
Pupils above the ACS (%)	82.8	81.1	95.2	77.5	83.2	87.2	83.8	80.0	75.7
Agreement of ratings & estimated mastery[b]	79.2	81.4	84.4	75.1	79.4	88.0	87.5	88.2	83.8
Number of students in the analyses	835	1016	895	724	712	741	767	797	1825
Number of teachers in the analyses	73	84	35	65	65	30	75	69	37

a) See ~~part~~ Section Variation and use of the cut-off scores.

b) See ~~part~~ Section Predictability of teacher ratings.

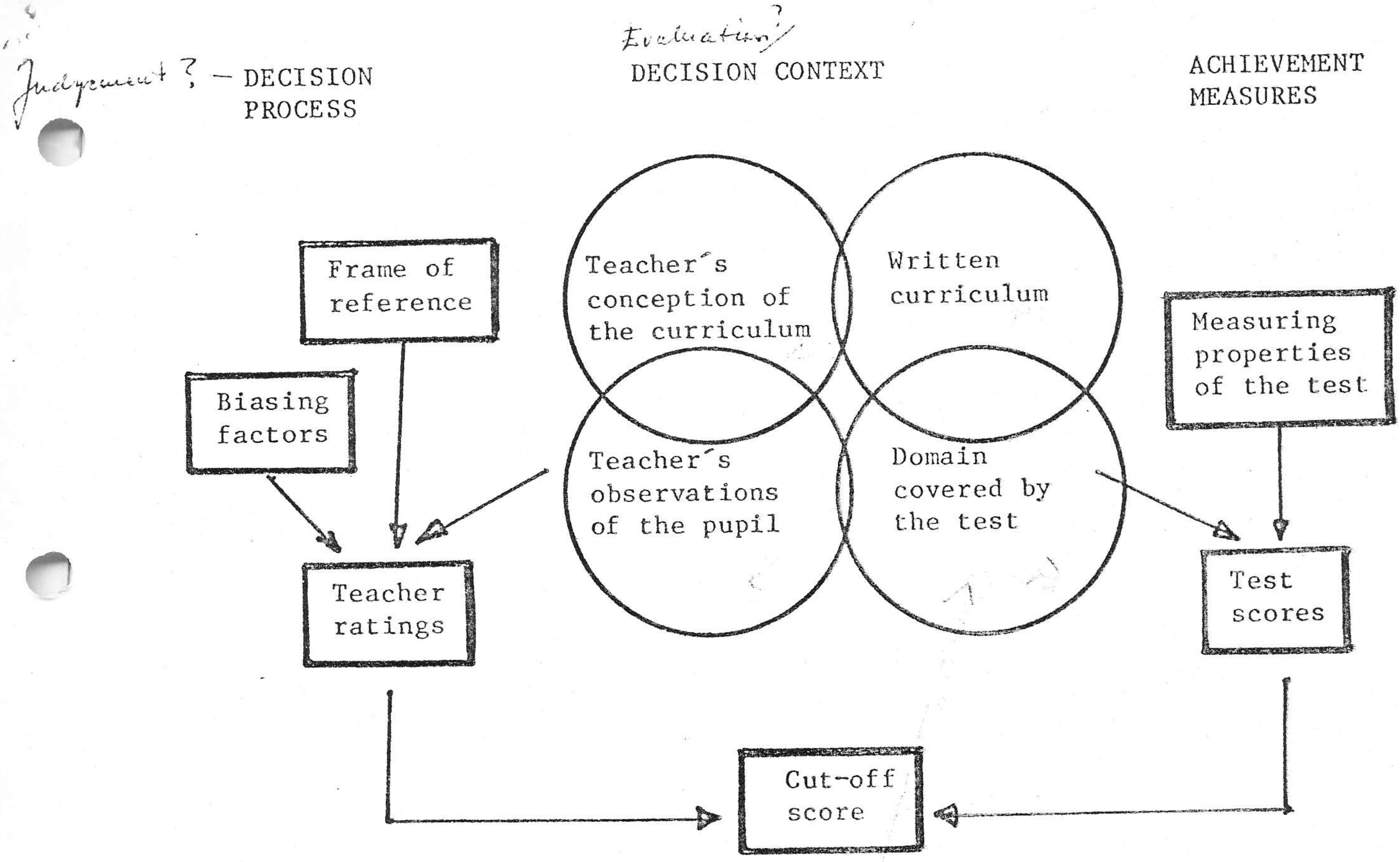
Judgement ? – DECISION PROCESS
Evaluation?
DECISION CONTEXT
ACHIEVEMENT MEASURES
Frame of reference
Biasing factors
Teacher ratings
Teacher´s conception of the curriculum
Written curriculum
Teacher´s observations of the pupil
Domain covered by the test
Measuring properties of the test
Test scores
Cut-off score

Figure 1. Factors affecting the cut-off scores in the Contrasting Groups method.

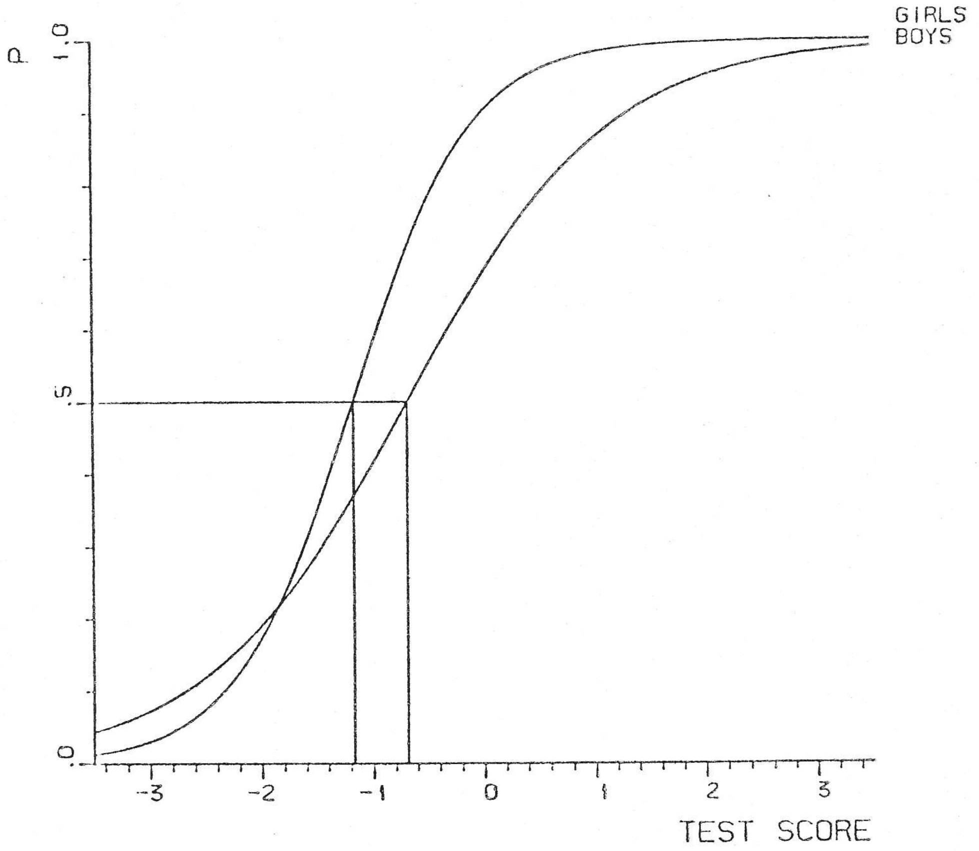

Figure 2. Probability of positive teacher rating (P) as a function of test score for boys and girls (class mean = 0) from the final model of Reading Comprehension - Finnish, 3rd grade.