It is argued that judgments in evaluative research are ultimately subjective, but that good criteria are available to assess their quality. One of these criteria is the robustness of the judgments against incompleteness or uncertainty in the data used to describe the educational system. The use of the robustness criterion is demonstrated through the case of a recent evaluation project in which the state of elementary education in The Netherlands was evaluated. To test robustness, four different procedures were simulated for item removal: (1) scaling; (2) removal of easy items; (3) removal of difficult items; and (4) removal of extreme items. The robustness study demonstrated that the qualifications used in the evaluation project were quite stable under the removal of items from the pool by these four methods. Nearly all the qualifications met the rigorous criterion of robustness. An appendix discusses the independence of the mean observed score of covariation between abilities. (Contains 3 tables, 8 figures, and 17 references.)
Robustness of Judgments in Evaluation Research

Wim J. van der Linden
Michel A. Zwarts

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente
Robustness of Judgments in Evaluation Research

Wim J. van der Linden
Michel A. Zwarts

To appear in Tijdschrift voor Onderwijsresearch
Robustness of Judgments

Abstract

The point of view is taken that judgments in evaluative research are ultimately subjective but that good criteria are available to assess their quality. One of these criteria is robustness of the judgments against incompleteness or uncertainty in the data used to describe the educational system. The use of the robustness criterion is demonstrated for the case of a recent evaluation project in which the state of elementary education in The Netherlands was evaluated.
Typically, the first stage of an evaluation project consists of a careful
description of the state of an educational object or system. In the next stage, the
state of the system is evaluated through a series of evaluative statements or
judgments. Examples of such judgments are: "The quality of teaching in the system
is excellent"; "Too many students in the system do not reach a satisfactory level of
proficiency in physics"; and "School management is poor". If the goal of the
evaluation project is to serve a reorientation of a policy with respect to the system,
the judgments usually result in a series of recommendations to improve the
functioning of the system.

For the descriptive stage, the standard methodology of empirical research
in the social sciences is available. This methodology includes the use of such
methods as survey and observation as well as various techniques of (multivariate)
descriptive statistical analysis to summarize the results. Though descriptive
statements can be founded on a rigorous methodology, judgments seem to lack this
support. The main reason is the use of such qualifications as "excellent", "not
satisfactory", and "poor" in the examples above. The choice of such qualifications,
as well as their definitions, is a subjective matter. However, subjectivity is not
necessarily erratic, and criteria for good qualifications do exist. Judgment does not
imply lack of rationality.

One criterion for the quality of judgments is consistency. For example,
suppose that empirical research has shown time and again that certain instructional
measures lead to an increase in the achievements of the students in a given domain,
and that a system to be evaluated scores high on the use of these measures. Then, ignoring the role of costs as well as the possibility of interaction between factors in the system, it seems inconsistent to make judgments that provide the former finding with a negative and the latter with a positive qualification. Such evaluations are inconsistent in the sense that they imply a world that can never exist. It should be noted that in this example empirical research was used to show that a set of qualifications is inconsistent. Empirical research can only provide the evaluator with objective information about what worlds are possible and what not. It remains a subjective choice to evaluate one possible world over the other.

Another obvious criterion is explicitness. The criterion of explicitness includes the requirement that all judgments be based on explicit definitions of the qualifications and procedures used in the evaluation. If this requirement is not met, the evaluator can never communicate his evaluations to others in a meaningful way. Also, it will never be possible to test these evaluations for consistency in the sense defined above.

It is not the purpose of this paper to give an extensive overview of criteria for the use of qualifications in evaluation research (for a more complete review, see van der Linden, to appear). Rather, the emphasis is on one criterion of a more technical nature than the previous examples. The criterion is necessary because judgments may have to be based on a description of the state of the system which is incomplete, uncertain, or erroneous due to the quality of the data. An example is an evaluation project in which the state of some relevant throughput factor is not precisely known. In such a case, which is certainly not untypical of educational evaluation, the evaluator may have to base his or her judgments on a best guess as to the state of this part of the system. An important criterion for the quality of his or her judgments, then, is robustness. Generally, a judgment is robust if minor
changes in the description of the state of the system do not lead to changes in the qualifications used in it. The idea underlying this criterion is obvious: Uncertainty about some part of the state of the system is less critical, the less dependent the qualifications are on the precise state the part of the system is in. The robustness of qualifications is usually assessed through a series of analyses in which changes in the values of some of the variables are made to simulate uncertainty about the state of the system, whereafter it is determined to what extent the qualifications would have to change. Obviously, robustness analyses are only possible if both the qualifications and the procedures leading to them are defined explicitly.

In the remainder of this paper, the results from a robustness study in a recent evaluation project in The Netherlands are reported to illustrate the possible contribution of robustness analysis to educational evaluation. The project was run by the Committee for the Evaluation of Elementary Education (CEB). In the next section, the problem addressed in the study is described. Subsequently, the methods of analysis will be given and the results will be discussed. The paper concludes with a discussion of the practical implications of the study.

Introduction to the Problem

The evaluation committee was appointed by the Dutch Secretary of Education in 1991. Its mission was to evaluate the state of elementary education in the Netherlands from 1988-1992. In particular, the interest was in an evaluation of four different aspects of elementary education in this period, its level of achievements being one of them. The results of the evaluation were published recently (Commissie Evaluatie Basisonderwijs, 1994a, 1994b, 1994c, 1994d, 1994e).
A fuller description of the assignment to the committee is given in Janssens (1995).

The committee had to report its findings at a level of aggregation that would suit a possible reorientation of the current policy of the Ministry of Education with respect to elementary education. Another constraint was that resources for data gathering were limited, and that the committee had to use existing sources of empirical data to perform its evaluation.

To present its evaluation of the achievements, the committee used the item material and scales from PPON. In this large-scale program for the assessment of educational progress in The Netherlands, which is run by the National Institute for Educational Measurement (Cito), the level of achievement in elementary education is periodically fathomed. The basic methodology used in PPON to scale the item pools and score the achievements is item response theory (IRT). The use of this methodology restricts the scaling of the items to the level of homogeneous subsets of the pool each measuring the same ability. An overview of the number of scales that were necessary to scale the item pools for the various subjects is given in Table 1.

For a complete review of the methodology used in PPON as well as reports of its assessments, the reader should consult van der Schoot (1993), Sijtstra (1992), Vinjé (1993), van Weerden (1993), Wijnstra (1998, 1990), and Zwarts (1990)
Definition of Qualifications

The selection of the qualifications by which the committee evaluated the achievements was guided by various considerations, three of which need to be explained here to be able to define the research problems addressed in this paper:

1. As already mentioned, the evaluation had to be reported at a level of aggregation suitable for recommendations on policy decisions. Therefore, it was necessary to combine sets of separate PPON scales into higher-level measures of achievement. For example, six separate scales for reading (Reading Reports; Reading Persuasive Texts; Reading Arguments; Reading References; and Reading Tables and Graphs) were combined into a single measure for Reading Comprehension. As IRT scales were not possible at this level of aggregation, the simple number of items correct score was used as a measure of achievement. However, this measure can be estimated from the scores on the IRT scales underlying the aggregate (see below). The number of aggregates in the evaluation is given in the last column of Table 1.

2. A second form of data reduction was also necessary to report the evaluations. The achievements of the population of students were in the form of distributions of scores. A usual way of defining qualifications for distinguishing between "good distributions" and "bad distributions" is in terms of their moments. Based on displays of the estimated distributions of the observed scores, the committee opted for qualifications for the first moments or means of the distributions. The main purpose of inspecting the displays was to get familiar with the relation between the location of the means and the shape of the left tails of the distribution. The qualifications were knowingly selected to be conservative; that is, relatively large
proportions of students in the population had to be at the lower ends of the achievement scales before an unfavorable qualification applied.

3. Instead of qualifications in the form of a simple good/bad dichotomy, the committee chose three different qualifications for which the terms "Satisfactory" (Dutch: voldoende), "Moderate" (Dutch: matig) and "Unsatisfactory" (Dutch: onvoldoende) were used. As a compromise between the fact that evaluations in terms of observed scores are dependent on item pool content and the fact that a single set of qualifications is easier to communicate, the committee opted for a common definition of qualifications with adjustments for item pools that were deemed to be too difficult or too easy.

In fact, the definition of the qualifications was a long process in which such factors as familiarity with the curriculum, teaching practices, quality of the learning materials, previous evaluations, and extensive consulting of relevant parties played an important role. The results are given in Table 2.

---

Table 2 about here

---

Estimation of Mean Observed Scores

Two typical distributions of observed scores are given in Figure 1. Both distributions were estimated using the assumption of a correlation equal to .80

---

Figure 1 about here
between the abilities on the underlying IRT scales.

The distribution for Calculating was evaluated as "Moderate". Its mean was just higher than the lower bound for this category but some 13% of the examinees solved less than one third of the items correctly. The distribution for Proportions/Percentages was estimated to have a mean in the category "Unsatisfactory". In this distribution, 36% of the examinees had less than one third of the items correct.

The means of the observed-score distributions were calculated from the item parameters estimated in the PPON projects. These estimates were obtained under the one-parameter logistic model with imputed values for the discrimination parameter (Verhelst, Glas & Verstralen, 1994). The ability distributions were scaled to be normal with mean 250 and standard deviation 50. Under the previous assumptions, the mean of an observed-score distribution can simply be calculated from the common marginal ability distribution and the sum of the response functions. This claim is proved in the Appendix.

Research Problems

The decision to use PPON item material and scales entailed two questions both related to the use of IRT in PPON.

First, though there is national agreement that the blueprints for the item pools had high content validity and that the sets of items in the pools covered the blueprints, some of the items were removed from the original pools in the scaling process. For example, for Arithmetic 4% of the items was removed from a pool of 491 items, whereas for Dutch 6% was removed from a pool of 498 items. These numbers are not large but important enough to pay attention to. As these items were removed on the basis of values of psychometric parameters and not of their content,
it seems safe to conclude that:

1. The resulting pools still define the same ability variables, and that these variables have therefore not lost their validity; and

2. The removal of some of the items from the pools may nevertheless have had effects on the observed-score distributions, and hence on the judgments by the committee.

An important question is how serious these possible effects are.

Second, only the marginal ability distributions were available from PPON. As already explained, the choice for the mean as the critical moment of the distribution of observed-scores was based on plots of observed-score distributions. However, under the assumption of multivariate normality, to be able to plot observed-score distributions for aggregates of IRT scales, Pearson’s correlation between the abilities must be known. (Remember that this requirement does not hold for the mean of the distributions.) As the abilities in each aggregate were “close”, and numerous research projects have shown high correlations between subtests covering different aspects of, for example, language and arithmetic, the assumptions of correlations in the neighborhood of .80 seemed realistic. An important question is how serious the consequences of violation of this assumption are.

Both questions were addressed in a robustness study.
Method

Removal of Items

Four different procedures of item removal were simulated. In each procedure, after the removal of an item the mean of the observed-score distribution was calculated, and the correct qualification from Table 2 was selected.

The following procedures were studied:

1. Scaling. The pair of items with the smallest difference between their values for the difficulty parameter was selected, and one item of the pair was chosen at random and removed from the pool. The mean of the observed-score distribution was calculated, and the appropriate qualification was identified. The steps were repeated until the pool was empty. This procedure simulates item analysis in which the range of the scale values of the items has to remain maximal but redundancies are removed by eliminating items from subsets that cluster too strongly. The procedure applies when the ideal is a pool of items with uniformly distributed scale values.

2. Easy Items. The item with the smallest value for the difficulty parameter was removed from the pool, the mean of the observed-score distribution was calculated, and the appropriate qualification was identified. The steps were repeated until the pool was empty. This procedure simulates the case where the item pool is considered too easy.

3. Difficult Items. The previous procedure was repeated, but now at each step the most difficult item was removed.

4. Extreme Items. This procedure is a combination of the previous two procedures. Alternately, the easiest and the most difficult item were removed. This procedure
simulates the case where the item pool is considered to be on target but, for example, the distribution of abilities of the examinees is expected to have less spread than the item pool.

**Correlation between Abilities**

To assess the robustness of the observed-score distributions with respect to the correlation between the abilities, a Monte Carlo method was used to generate observed-score distributions on the sets of items in the aggregates for various values of the correlation coefficient. As a correlation between the abilities lower than .60 was most unlikely, the following values for the correlation coefficient were used: .60, .70, .80, and .90.

In the description of the Monte Carlo procedure below, the notation of the variables is the same as in the Appendix but the indices \( j = 1, \ldots, J \) and \( i = 1, \ldots, I \) are now used to denote the abilities and the items in a subset for the same ability, respectively:

1. For each simulated examinee, the values of the vector of abilities \((\theta_1, \ldots, \theta_j)\) were drawn from a multivariate normal distribution with the assumed (common) value of the correlation coefficient.
2. The true scores \((t_1, \ldots, t_j)\) were calculated as

\[
t_j = \sum_{i=1}^{I} P_{ij}(\theta_j), \quad j = 1, \ldots, J, \tag{1}
\]

and normed on \([0, 1]\).
3. The conditional distributions of \(X_{ij}\) given \(T_j = t_j\) are generalized binomial. Their probability functions, \(\text{Prob}(X_{ij})\), were calculated using the first term in the expansion of the generalized binomial probability function given in...
Robustness of Judgments

4. The probabilities of the number-correct scores, \( \sum_{j=1}^{J} X_j \), were calculated as

\[
Prob(T=t) = \frac{\sum X_j = t}{\prod_{j} \text{Prob}(X_j)}.
\] (2)

The last step in the procedure made use of the fact that for a fixed examinee the observed scores \( X_j \), \( j=1,...,J \), were independent.

The accuracy of the approximation in Step 3 was checked against an algorithm suggested by Lord and Wingersky (1984) which produces the full generalized binomial distribution (see below).

The procedure was repeated for \( N=10,000 \) examinees. It should be noted, however, that for each examinee not one realization of \( X_j \) given \( T_j=t_j \) but its full conditional distribution was generated. The number is thus large enough to guarantee a smooth and stable result.

**Results**

Graphs are used to present the results for the scaling procedure. In Figure 2, the mean observed relative scores for the five aggregates in Arithmetic are displayed as a function of the proportion of items removed due to scaling.
qualifications defined in Table 2. Generally, the curves follow a flat course, indicating extreme robustness of the mean with respect to the removal of items due to scaling. To cross one of the lines, the removal of 91% of the items for Basic Skills and 100% of the items for Proportions/Percentages was needed. For Calculating, the percentage was equal to 62%. The percentages for Fractions and Measurement are lower but still equal an impressive 45% and 33%, respectively. After these values, the two last curves started moving back and forth between the two sides of the upper (Fractions) and lower lines (Measurement). This behavior is typical of mean scores that were close to the borderline between two qualifications, remained there after removal of the items, but showed small fluctuations.

The results for the aggregates in the other subjects are given in Figures 3 through 6.

The results are generally the same as for Arithmetic. All curves had a flat course, and, except for Reading English, at least 30-40% of the items had to be removed before the qualifications change. The case of Reading English is an interesting one. The curve was flattest of all curves in Figures 2-6, but the curve coincided with the upper line nearly perfectly. The same phenomenon was observed for Reading Comprehension. Its curve was also flat and uniformly close to the line between "Satisfactory" and "Moderate". Nevertheless, 38% of the items had to be removed from the pool to change the qualification. At a later stage, the curve moved back to the original qualification. In its report, the committee made the provision that important parts of this aggregate were less favorable than the general impression.
suggested. Also, uncertainty was expressed due to the fact that data from an international comparison of achievements in Reading Comprehension had yielded conflicting information (Commissie Evaluatie Basisonderwijs, 1994a, sect. 5.1).

The results for all four principles of item removal are given in Table 3. The first column gives the percentages of items that had to be removed for the scaling procedure. The next three columns present the results for the other item removal procedures. Obviously, removal of the most difficult or easy items introduced a shift in the observed-score distributions, and generally the qualifications changed earlier than in the previous case. Nevertheless, with the exception of Measurement and Reading Comprehension for the removal of the easiest items and Biology for the most difficult items, the qualifications were remarkably robust for all aggregates. In these exceptional cases of change, again the mean observed scores were already close to the borderline between two classifications for the intact item pool. For example, for Reading Comprehension the mean relative observed score for the intact item pool for the pool was .71, a result close to the cut-off score of .70 separating "Satisfactory" from "Moderate" (see Figure 2). The removal of the items with extreme difficulty values at both ends of the scale had, except for Reading of English, no noticeable effect on the qualifications. In the majority of the cases, nearly all items had to be removed before the qualification changed.
In Figure 7, two typical observed-score distributions for values of the correlation coefficient in the range from .60-.90 are shown. The effect of lowering the value of the correlation was a small shift of the mode of the distribution to the center of the scale. (However, remember that this phenomenon does not hold for the mean of the distribution. This parameter is independent of the value of the correlation coefficient.) Consequently, the value of the correlation coefficient does have some effect on the left tail of the distribution, but the effect is not dramatic. It seems safe to conclude that the relation between the mean and the left tail of the distributions observed by the committee does not change much in the neighborhood of $r=.80$.

As already observed, in Step 3 of the procedure for generating the observed-score distributions, an approximation to the generalized binomial distribution of $X$ given $T=t$ was made. The quality of the approximation was checked by comparing its results against those obtained for the exact distributions using the computer program AAPMOMT which implements the algorithm by Lord and Wingersky (1984) referred to earlier. The results were always virtually identical. Figure 8 gives the distributions for the same two aggregates as in Figure 7.
Robustness of Judgments

16

The approximation proved to be excellent; the difference between the results of the two methods is hardly discernible.

Discussion

The main conclusion from the robustness study reported in this paper is that the qualifications used in the evaluation project are quite stable under the removal of items from the pool according to the four procedures defined above. Nearly all of the qualifications thus met a rigorous criterion of robustness.

In this study, the results for the scaling procedure are most important since this procedure comes closest to the procedure actually used in the PPON projects. However, it should be noted that the former is an idealized version of the latter, and that differences between the two may exist. Also, the procedure was applied to the item pools that were the results from PPON item analyses, and not to the original pools. Generalizing the findings to the original pools thus involves an element of extrapolation, albeit that the differences between the sizes of the two kinds of pools were generally small. Also, the fact that, with a few exceptions, remarkably robust results were obtained for procedures that deliberately made the item pools easier or more difficult does lend some support to the claim that this generalization is unlikely to involve serious bias.

It is emphasized that robustness of qualifications is only one necessary criterion which judgments in evaluation projects must meet, and that judgments are not automatically meaningful if they are robust. However, as illustrated in this paper, if uncertainty exists as to the knowledge base on which the judgments have to based, then robustness analysis is an excellent means to assess how serious the consequences of this uncertainty are.
Appendix

Independence of Mean Observed Score of Covariation between Abilities

For ease of exposition, the case of two distinct abilities is addressed. Let \( \theta_1 \) and \( \theta_2 \) be these two abilities. The bivariate distribution of the two abilities is represented by probability density function \( f(\theta_1, \theta_2) \), whereas the marginal distributions of \( \theta_1 \) and \( \theta_2 \) are denoted as \( f_1(\theta_1) \) and \( f_2(\theta_2) \). Let \( X_1 \) and \( X_2 \) be the observed scores on the item sets measuring \( \theta_1 \) and \( \theta_2 \) and \( T_1 \) en \( T_2 \) the classical true scores for these observed scores.

In PPON, the marginal distributions of \( \theta_1 \) and \( \theta_2 \) are scaled to have common marginal densities:

\[
f_1(\theta_1) = f_2(\theta_2) = f(\theta).
\]

This feature is used in the proof below. The first step in the derivation follows from classical test theory, whereas the other steps are straightforward. Indices i and j denote items measuring the first and second ability, respectively. The proof runs as follows:
Robustness of Judgments

\[ E(X_1 + X_2) = E(T_1 + T_2) \]

\[
= \int \left[ \sum_i P_i(\theta_1) + \sum_j P_j(\theta_1) \right] f(\theta_1, \theta_2) d\theta_1 d\theta_2
\]

\[
= \int \sum_i P_i(\theta_1) f(\theta_1, \theta_2) d\theta_1 d\theta_2 + \int \sum_j P_j(\theta_2) f(\theta_1, \theta_2) d\theta_1 d\theta_2
\]

\[
= \int \sum_i P_i(\theta_1) f_1(\theta_1) d\theta_1 + \int \sum_j P_j(\theta_2) f_2(\theta_2) d\theta_2
\]

\[
= \int \left[ \sum_i P_i(\theta) + \sum_j P_j(\theta) \right] f(\theta) d\theta.
\]

Hence, when calculating the mean observed score, possible covariation between the underlying abilities can be ignored, and the item response function may be summed across abilities.
References


### Table 1.

**Aggregation of PPON scales in evaluation project**

<table>
<thead>
<tr>
<th>Subject</th>
<th># Original Scales</th>
<th># Aggregates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch Language</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>World Orientation</td>
<td>30</td>
<td>8</td>
</tr>
<tr>
<td>English</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Traffic</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. World Orientation is a combination of subjects. See Table 3.*
Table 2.

Definition of qualifications used in evaluation

<table>
<thead>
<tr>
<th>Qualification</th>
<th>Mean of Score Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfactory</td>
<td>&gt; 70%</td>
</tr>
<tr>
<td>Moderate</td>
<td>55% - 70%</td>
</tr>
<tr>
<td>Unsatisfactory</td>
<td>&lt; 55%</td>
</tr>
</tbody>
</table>

Note. For item pools judged to be too difficult a downward adjustment of 10% and 5% was made for the lower bounds of Satisfactory and Moderate, respectively.
Table 3

Percentages of items needed to change the qualifications for the four methods

<table>
<thead>
<tr>
<th>Subject</th>
<th>Scaling</th>
<th>Easy</th>
<th>Difficult</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic Skills</td>
<td>91</td>
<td>14</td>
<td>27</td>
<td>100</td>
</tr>
<tr>
<td>Calculating</td>
<td>62</td>
<td>29</td>
<td>16</td>
<td>100</td>
</tr>
<tr>
<td>Fractions</td>
<td>45</td>
<td>17</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Proportions/Percentages</td>
<td>100</td>
<td>100</td>
<td>14</td>
<td>100</td>
</tr>
<tr>
<td>Measurement</td>
<td>33</td>
<td>3</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>37</td>
<td>3</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Listening</td>
<td>94</td>
<td>32</td>
<td>100</td>
<td>91</td>
</tr>
<tr>
<td>Composition</td>
<td>71</td>
<td>21</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Spelling</td>
<td>39</td>
<td>34</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Grammar</td>
<td>79</td>
<td>54</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Parsing</td>
<td>71</td>
<td>37</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Language Reflection</td>
<td>100</td>
<td>32</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>World Orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>41</td>
<td>28</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>Physics</td>
<td>66</td>
<td>13</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>Regional Geography</td>
<td>88</td>
<td>26</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>Physical Geography</td>
<td>91</td>
<td>16</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>Topography</td>
<td>100</td>
<td>100</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>History</td>
<td>40</td>
<td>47</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Spiritual &amp; Religious Movements</td>
<td>78</td>
<td>30</td>
<td>100</td>
<td>39</td>
</tr>
<tr>
<td>Social Relations</td>
<td>97</td>
<td>20</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>3</td>
<td>3</td>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>Listening</td>
<td>96</td>
<td>41</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Speaking</td>
<td>97</td>
<td>27</td>
<td>14</td>
<td>100</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>59</td>
<td>24</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Use of Dictionary</td>
<td>100</td>
<td>75</td>
<td>100</td>
<td>55</td>
</tr>
<tr>
<td>Traffic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practical Skills</td>
<td>91</td>
<td>42</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Authors' Note

Both authors are equally responsible for the contents of this paper; the order of their names is alphabetical. Michel A. Zwarts is at the Inspectorale of Education, De Meern, Netherlands. The empirical results in this paper were presented earlier at a Vereniging voor Onderwijsresearch, Divisie Methodologie en Evaluatie, symposium, Arnhem, March 23, 1994.
Figure Captions

Figure 1. Estimated observed-score distributions for: (a) Calculating; and (b) Proportions/Percentages.

Figure 2. Mean observed score as a function of the proportions of items removed due to scaling for Arithmetic.

Figure 3. Mean observed score as a function of the proportions of items removed due to scaling for Dutch.

Figure 4. Mean observed score as a function of the proportions of items removed due to scaling for World Orientation.

Figure 5. Mean observed score as a function of the proportions of items removed due to scaling for English.

Figure 6. Mean observed score as a function of the proportions of items removed due to scaling for Traffic.

Figure 7. Estimated observed-score distributions for: (a) Calculating; and (b) Proportions/Percentages (different correlation between abilities).

Figure 8. Observed-score distributions estimated by: (a) Taylor approximation to generalized binomial; and (b) exact distribution function.
Robustness of Judgments

Calculating

Proportions/Percentages

\[
\begin{align*}
\text{Density} & \quad 0 & \quad 50 & \quad 100 & \quad 150 & \quad 200 & \quad 250 \\
\text{Relative score} & \quad 0.0 & \quad 0.2 & \quad 0.4 & \quad 0.6 & \quad 0.8 & \quad 1.0
\end{align*}
\]

\[
\begin{align*}
\text{Density} & \quad 0 & \quad 50 & \quad 100 & \quad 150 & \quad 200
\end{align*}
\]

\[
\begin{align*}
\text{Relative score} & \quad 0.0 & \quad 0.2 & \quad 0.4 & \quad 0.6 & \quad 0.8 & \quad 1.0
\end{align*}
\]
Robustness of Judgments

Basic Skills

Calculating

Fractions

Proportions/Percentages

Measurement
Robustness of Judgments

BEST COPY AVAILABLE
Robustness of Judgments

Proportion items detected

Mean relative score

Proportion items deleted

Regional Geography

Physical Geography

Topography

Spiritual & Religious Movements

Social relations

BEST COPY AVAILABLE
Robustness of Judgments

30

**English Reading**

![Graph showing mean relative score vs proportion of items deleted for English reading.]

**English Listening**

![Graph showing mean relative score vs proportion of items deleted for English listening.]

**English Speaking**

![Graph showing mean relative score vs proportion of items deleted for English speaking.]

**English Vocabulary**

![Graph showing mean relative score vs proportion of items deleted for English vocabulary.]

**English Use of Dictionary**

![Graph showing mean relative score vs proportion of items deleted for English use of dictionary.]

*BEST COPY AVAILABLE*
Robustness of Judgments

Calculating

Proportions/Percentages
Calculating

Proportions/Percentages

Robustness of Judgments
Titles of recent Research Reports from the Department of Educational Measurement and Data Analysis.
University of Twente, Enschede, The Netherlands.

RR-94-10 W.J. van der Linden & M.A. Zwarts, Robustness of judgments in evaluation research
RR-94-9  L.M.W. Akkermans, Monte Carlo estimation of the conditional Rasch model
RR-94-8  R.R. Meijer & K. Sijsma, Detection of aberrant item score patterns: A review of recent developments
RR-94-7 W.J. van der Linden & R.M. Luecht, An optimization model for test assembly to match observed-score distributions
RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, Some new item selection criteria for adaptive testing
RR-94-5 R.R. Meijer, K. Sijsma & I.W. Molenaar, Reliability estimation for single dichotomous items
RR-94-3 W.J. van der Linden, A conceptual analysis of standard setting in large-scale assessments
RR-94-2 W.J. van der Linden & H.J. Vos, A compensatory approach to optimal selection with mastery scores
RR-94-1 R.R. Meijer, The influence of the presence of deviant item score patterns on the power of a person-fit statistic
RR-93-1 P. Westers & H. Kelderman, Generalizations of the Solution-Error Response-Error Model
RR-91-1 H. Kelderman, Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory
RR-90-7 E. Boekkooi-Timminga, A Method for Designing IRT-based Item Banks
RR-90-6 J.J. Adema, The Construction of Weakly Parallel Tests by Mathematical Programming
RR-90-5 J.J. Adema, A Revised Simplex Method for Test Construction Problems
RR-90-4 J.J. Adema, Methods and Models for the Construction of Weakly Parallel Tests
RR-90-2 H. Tobi, Item Response Theory at subject- and group-level
RR-90-1 P. Westers & H. Kelderman, Differential item functioning in multiple choice items

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.