ED 388 683 TM 023 717

AUTHOR Kuhlemeir, Hans; And Others

TITLE Multilevel Factor Analysis Applied to National Assessment Data.

PUB DATE Apr 95
NOTE 37p.; Paper presented at the Annual Meeting of thc American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; *Factor Analysis; Factor Structure; Foreign Countries; High Schools; *High School Students; Language Proficiency; *National Competency Tests; National Surveys; School Effectiveness; Test Reliability; *Test Use; Writing (Composition); Writing Evaluation; *Writing Tests
IDENTIFIERS Dimensionality (Tests); *Multilevel Analysis; Netherlands; *Performance Based Evaluation

ABSTRACT

In the Dutch Educational Assessment Program, the students' language proficiency is measured in grade 9, at age 15. Writing performance is measured through several performance-based writing tasks, rated on numerous aspects such as content, style, organization, punctuation, spelling and grammar. As a consequence, national performance levels are reported in a rather detailed fashion. This paper focuses on the efficiency of the measurement and reporting of writing performance using a sample of 1,451 students during 1987-88. Multilevel factor analysis (MLFA) was used to determine dimensionality and reliability. At the school level, a distinction between two factors was warranted: composition versus technical conventions. At the student level, the composition factor remained intact, but the technical conventions factor had to be divided into the subskill factors of punctuation, spelling, and a factor interpreted as technical quality at the sentence level. Factor structure was not found to be independent of the writing task. Indications of the unreliability of measurement of functional writing proficiency suggest that the usefulness of the national assessment data for school effectiveness studies is limited. Recommendations are made for improvement of assessment. (Contains 5 tables and 57 references.) (Author/SLD)

Multilevel factor analysis applied to national assessment data[1]

Hans Kuhlemeier[2], Huub van den Bergh[3] and Johan Wijnstra[2]

[2]. National Institute of Educational Measurement (Cito),
Department of National Assessment, Nieuwe Oeverstraat 65,
6801 MG Arnhem, Netherlands

[3] Utrecht University, Centre for Language and Communication,
Trans 10, 3512 JK Utrecht, Netherlands

[1] We are grateful to H. Goldstein and G.D. Mellenbergh for their constructive and insightful comments on an earlier version of this manuscript.

Summary

In the Dutch Educational Assessment Program, the students' language proficiency is measured in grade 9 (at age 15). Their writing performance is measured by means of several, performance-based writing tasks. Each task is rated on numerous aspects such as content, style, organisation, punctuation, spelling and grammar. As a consequence, national performance levels are reported in a rather detailed fashion. This article focuses on the efficiency of the measurement and the reporting of writing performance. Multilevel factor analysis (MLFA) was used to determine dimensionality and reliability. It appeared that the between and within-school factor structures were not identical. At the school level, a distinction between tw factors was warranted: composition versus technical conventions. At the student level, the composition factor remained intact. However, the technical conventions factor had to be divided into three subskill factors: punctuation, spelling and a factor that was interpreted as the technical quality at the sentence level. Furthermore, it was found that the factor structure was not independent of the writing task. It is also shown that, in general, functional writing proficiency is measured relatively unreliable, at the school, student, and score level. It is concluded that the potential usefulness of national assessment data for the purpose of school effectiveness studies is limited. Recommendations are made concerning the instrumentation of writing performance, the sample and administration design, the statistical analysis and the reporting of national assessment data.

Introduction

The measurement of writing performance has been an essential element in the recent language assessments in the Dutch language region (Wesdorp et al, 1986; Kuhlemeier & Van den Bergh, 1989; De Glopper, 1988; Vergeer & Oostdam, 1988; Zwarts, 1990; Rymenans, Leroy & Daems, 1991; Sytstra, 1993) and in other areas (Conry & Jerolski, 1980; Gorman, White, Orchard, Tate & Sexton, 1981; Neville, 1988; Applebee, Langer, Mullis & Jenkins, 1990). In national assessment, the knowledge and skills of the students are described in a comprehensive and detailed fashion (Bock, Mislevy & Woodson, 1982). The section on writing performance consists of multiple tasks, such as writing essays, letters and forms, since in the course of their writing education, students must learn to write various types of texts. Several studies have shown that the degree of across-task generalizability of performance-based writing assessment is limited (Quellmalz, Capell & Chou, 1982; Breland, Camp, Jones, Morris & Rock, 1987; Ackerman & Smith, 1988; Van den Bergh, 1988; Kuhlemeier & Van den Bergh, 1991; Schoonen, 1991; Rijlaarsdam, Van den Bergh & Zwarts, 1992; Linn & Burton, 1994). Writing performance depends on the task given. The task specificity is another important reason for including varied writing tasks in national assessment.

In the language assessments, the writing products of the students are rated on a large number of aspects by trained juries. Written texts must, after all, satisfy a number of different criteria not only in the realm of content, style, and organisation, but also such technical aspects as punctuation, spelling, and grammar.

The large number of tasks and aspects makes it possible to provide a comprehensive and detailed description of the quality of writing education. Besides the obvious psychometric advantages (see Beaton, 1988; Wijnstra, 1988), there are, however, some disadvantages. One of these is the amount of work involved in collecting, processing, analyzing and reporting on the performance on the many tasks with their numerous aspects. Another disadvantage is related to the (in)accessibility of the often large and detailed reports that have to be prepared, in particular, for non-researchers (Gipps & Goldstein, 1983). Often, the number of report units is as large as the number of writing tasks multiplied by the number of aspects.[1] It is debatable to what degree this can still be defined as efficient. In order to gain more insight into this problem, further research into the dimensionality of the measurement of writing performance is necessary.

In this study, the issue of efficiency in the measurement of writing performance in national assessment is dealt with. The data are taken from the National Assessment of Language Performance (Kuhlemeier & Van den Bergh, 1989). In this survey ninth-grade students (of app. the age of 15) in secondary education were given a number of different writing assignments. Each student did one or two of the assignments. Every writing product was rated on twelve different aspects by a trained jury. For the purpose of illustration, only two assignments and ten of the aspects were used in this article.

The main question pertains to the extent the two tasks and the ten aspects can be distinguished in the data. If the given tasks and aspects are in fact highly related, then these distinctions are meaningless from an empirical point of view. In that case, it would be sufficient to report on only one task and/or fewer aspects. The practical implication would be simpler measurement and a less differentiated report on the state of the art of writing performance.

The question of the dimensionality of the measurement also plays a role at the school level. One of the goals of national assessment in the Netherlands is to assist schools in evaluating their effectiveness. Although national assessment has neglected the school as a unit of analysis, reports of the school means and their distribution are at least as important as the tradi'ional reports per subpopulation (see Bock, 1988, p. 748). Accordingly, in a recent assessment (Kuhlemeier et al., 1994) each school received a so-called school report. This report provides, among other things, the scale means on a large number of curricular elements. It enables teachers and principals to evaluate the effectiveness of their school in comparison to a national sample of schools working in similar circumstances (intake, school type). Perhaps this rather detailed report could have been simplified. If the measurement had been unidimensional at the school level, than a much simpler school report would have been possible.

Another goal of national assessment is the provision of indications for fundamental scientific (in-depth) studies and evaluation research. Currently, research into the effectiveness of schools in the Netherlands is high on the research agenda. Scheerens & Stoel (1987) proposed incorporating part of the school effectiveness research into the national assessment program (p. 12). If, on the basis of the regular assessments, it is possible to differentiate between high- and low-performance schools, then this could be a reason for follow-up investigations at schools which participated in the assessment (outlier studies, for example, in which high and low achieving schools can be compared on the basis of specific characteristics). If the measurement of writing performance would be unidimensional, than only one single indicator of the mean achievement level of the school would be necessary.

Multilevel factor analysis

Until recently the dimensionality of writing performance was analyzed almost exclusively using conventional unilevel analysis techniques such as explorative or confirmative factor analysis via covariance structure analysis (Quellmalz, Capell & Chou, 1982; Breland et al., 1987; Van den Bergh, 1988; Kuhlemeier & Van den Bergh, 1991; Sytstra & Zwarts, 1991; Zwarts & Rijlaarsdam, 1991). These techniques are based on assumptions, some of which are not always fulfilled. Two of these are of particular importance.

The first assumption is of a statistical nature. Classical factor analysis assumes that the data are collected as a random sample from the population and that the observations are statistically independent. In national assessment, a multi-stage sampling procedure is implemented. First schools are drawn, for example, and then students within the schools are selected. Because randomly selected students from one school are more alike than randomly selected students from different schools (due to selection and instruction), the assumption of statistical independency is violated. In that case, the precision of a cluster sample is lower than that of a random sample with the same number of observations (Kish, 1965). All other things being equal, the precision decreases as the intraclass correlation increases (Tate & King, 1994). In a two-stage sample with students within schools, the intraclass correlation is defined as the proportion of the total variance that lies between schools (Kish, 1965). Although multiple matrix sampling designs (Sirotnik, 1974) are used to minimize cluster effects, significant intraclass correlations often do occur (Johnson & Rust, 1992). Results from national assessments in the U.S. and the Dutch language area show that the intraclass correlations usually varies between .1 and .5 (De Glopper, 1988; Kuhlemeier & Van den Bergh, 1989; Rymenans, Leroy & Daems, 1991; Tate & King, 1994). These effects are clearly significant and meaningful (Blok & Eiting, 1988). This inevitably leads to the conclusion that the observations are not independent from a statistical point of view. In that case unilevel analysis overestimates the precision and leads too quickly to significance of the analyzed relations (Goldstein, 1987).

A second assumption is of a substantive nature. In classical factor analysis, it is assumed that the factor structure is invariant across educational levels. This is by no means an absolute rule. Sometimes a factor can be demonstrated at the student level, but is not demonstrable at the higher level of the school (Longford, 1990; Balke, 1992). There are systematic differences between students (within schools), but the mean level does not vary from school to school. Conceivably, the number of factors as well as the correlations between the factors could vary from level to level (Robinson, 1950). Melse & Kuhlemeier (1993) studied the structure of German-language performance using a three-level factor analysis (scores nested within students, who are nested within classes). At the class level, the same language factors were shown to be relevant as were relevant at the student level. The factor variances and covariances at the class level were, however, different from those at the student level. Hox (1993) analyzed the scores of 187 children from 37 large families on the GIT (Groninger Intelligence Test) by means of a two-level factor analysis (children nested within families). Different structures were found at the between- and within-family level. At the lowest level (children within families) two factors were necessary in order to describe the correlations between the six subtests; at the higher level, however, these two factors correlated perfectly. The between-family structure was interpreted by Hox as the common influence of heredity and environmental factors; the within-family structure was assumed to represent the individual, idiosyncratic contribution of the child within his or her family. In school effectiveness research, the interpretation of between- and within-school factor structures is a rather unexplored area. It seems, however, plausible to interpret the within-school structure as the cognitive or mental structure in students, analogous to the interpretation of a classical factor analysis. The between-school structure refers to differences between schools in the selection of students, the curricular offering and the effectiveness of instruction and school policies (Longford & Muthén, 1992; Melse & Kuhlemeier, 1993). It is this between-schools structure that is of major importance in studying school effectiveness.

Given these statistical and theoretical arguments a multilevel factor analysis (MLFA) was chosen for use in this study (Goldstein & McDonald, 1988). Using this technique, it is possible to analyze the factor structure at multiple levels simultaneously (Longford, 1990; Muthén, 1990; Raudenbush, Rowan & Kang, 1991; Longford & Muthén, 1992). It also provides more accurate estimates of the standard errors (Goldstein, 1987). In addressing the question of the dimensionality and the hierarchical structure of national assessment data, five sub-questions are distinguished:

1) How many different factors have to be distinguished at the school and the student level in order to provide an adequate description of the writing performance of ninth graders in secondary education?

2) How large are the differences between schools in the mean levels of writing performance?

3) How large are the differences between students?

4) What is the correlation between the factors at the school and student level?

5) How reliable are measurements of writing performance at the school and student level?

Research methods

Subjects

The data were taken from the 1987-1988 National Assessment of Language Performance in the Netherlands (Kuhlemeier & Van den Bergh, 1989). In this study, a nationally representative description is given of Dutch language achievement in the third year of secondary education. The data were collected at the end of the 1987-1988 school year. In this article, only those scores were used which the 1451 students were given for two writing assignments: a letter of apology and a letter of application. The students were drawn from 184 schools with the following breakdown for school type or track: pre-university education (19%), higher general secondary education (19%), intermediate general secondary education (24%), junior vocational education (21%) and junior vocational domestic-science education (18%); 48% were boys and 52% were girls; the mean age of the students was approximately 15 years and 6 months.

Instruments

The letter of apology and the letter of application were functional writing tasks. The first assignment, the letter of apology, consists of writing a letter to the principal at the school. The student has to apologize for missing an appointment owing to illness. The letter of application is addressed to the head of the personnel department of a large department store, in response to an advertisement asking for a summer-season worker. Both letters are 'specified' tasks, the most frequently used type in national assessment (Schoonen, 1991). Here, as opposed to the traditional essay assignment, the communicative context - goal and audience - is defined. As in the case of the essay, however, the student does have freedom of choice in determining content, style, organisation, and the more technical aspects of the text. The two writing tasks are very similar. They are both examples of short, functional letters. Therefore they may be regarded as two equivalent measures of functional writing ability or as samples from approximately the same content domain.

The letters were rated on ten different aspects, that is, text characteristics. A priori these can be divided into two clusters: a composition skills component (composition) and a technical writing skills component (technical conventions).

With regard to the composition component five ratings are available for both the letter of apology and the letter of application:

1) Global quality. The jury had to render an initial general impression after a single reading of the text.

2) General content. In assessing this aspect, the jury had to render a general opinion on the quality of the content of the letter.

3) Style. What was assessed here was whether the use of language in the letter was appropriate and comprehensible and did not interfere with communication; it also had to be taken into consideration whether :he text was vague or to the point.

4) Organisation. The assessment of organisation dealt with the manner in which students presented the text. Muddled or poorly organized texts received low scores.

5) Content elaboration. In judging this characteristic, the jury used a checklist containing a number of relevant content elements.

With regard to the technical-ccnventions component, data on five types of writing errors are available:

6) Punctuation: the proportion of punctuation errors per hundred sentences.

7) Spelling: the proportion of spelling errors per hundred words.

8) Spelling of verbs: the proportion of spelling errors in verbs per hundred sentences.

9) Grammar and idiom: the proportion of grammatical and idiomatic errors per hundred sentences.

10) Sentence construction: the proportion of grammatically incorrect sentences per hundred sentences.

Each letter was rated on each aspect by a random sample of two or three of nine teachers. Prior to the rating of the writing products of the students, the teachers underwent a short but intensive training (Kuhlemeier & Van den Bergh, 1989). The aspects of global quality, content, style, and organisation were rated by three raters on a five-point scale. As an aid they used previously-scaled anchor texts which indicated texts of low, medium, and high quality. Of course, there were different anchor texts for different aspects. Two raters rated the remaining six aspects. Interrater reliabilities varied from .66 for the number of verb-spelling errors counted in the letter of apology to .99 for assessments of the content of the letter of application, with an average of .83 (Kuhlemeier & Van den Bergh, 1989). Additional descriptive statistics are given in the Appendix.

Structure of the data

The total of almost 70,000 scores given by the raters were summed up into individual student scores for each task and each aspect. In total, there are 28,635 of these student scores for 1451 ninth graders from 184 secondary schools. The number of students from each school varies from five to nine, with an average of 7.91. The maximum number of scores per student is twenty: ten for the letter of apology and ten for the letter of application. The actual number of scores per student varies from four to twenty, with an average of 19.73. The dataset is, therefore, almost complete.

Analysis

Consistent with the structure of the data, three levels are distinguished in the multilevel factor analysis: school (level 3), student (level 2) and score (level 1). Factor variation is decomposed into three parts:

1) a between-schools component (school level);
2) a between-students-within-schools level (student level);
3) a between-scores-within-student component (score level).

In this variance-component model, the variances and covariances of the factors are modelled at the student and the school level. The residual variance, which cannot be accounted for by these factors, is represented at the score level (Raudenbush, Rowan & Kang, 1991). The first level is the measurement level; it describes the relation between the observed scores and the factors in the structural part of the model, the between- and within-school level (level 2 and 3).

The scores are standardized for each task by aspect with a mean of zero and a variance of one. This standardization equates the means and the (total) variances, but does not, of course, equate the distribution of variances and covariances across different levels.

In the analysis three models were distinguished that differed only in the structure at levels 2 and 3; they share the same level 1 structure. At this intra-student-between-scores level, a separate residual variance was estimated for each combination of task and aspect. The degree to which the scores indicate the factors can, therefore, differ from task to task and from aspect to aspect. The proportion of residual variance at the first level provides insight into the 'uniqueness' of the factors at the second and third levels. In other words, the residual variance represents that part of the total variance which cannot be attributed to these factors and which, given the model, can be interpreted as error variance. Hence, the factors at the within- and between-school level are corrected for attenuation (Raudenbush, Rowan & Kang, 1991).

Basic model (Model I)

Initially, the structure of the data was described using a ten-aspect model (Model I). In this basic model, there is a separate factor (both at the school and the student level) for every ten of the aspects. Each aspect factor has two indicators: the scores on both tasks. This seems reasonable, since the two functional writing tasks are very similar. Suppose Y_{hijk} refers to the score on aspect h ($h = 1, 2, \ldots ,$

10), of task *i* (*i = 1, 2*), of student *j* (*j = 1, 2, ..., N_k*), in school *k* (*k = 1, 2, ..., N*), and X_{hijk} is a dummy-variable that is turned on (= 1) if it represents the score on aspect *h* of task *i*. In all other cases this dummy-variable is turned off (= 0). Now the model can be written as:

$$Y_{hijk} = \Sigma_h \; (ß_{h0jk} * X_{hijk}) + e_{hijk}. \qquad (1)$$

In Equation (1), $ß_{h0jk}$ is the regression weight for aspect *h* of student *j* in school *k*. $ß_{h0jk}$ can be interpreted as the mean score over both tasks of this student for aspect *h*. The second term, e_{hijk}, refers to the deviation for aspect *h* of task *i* from the mean score for this aspect of student *j* in school *k*. It is assumed that e_{hijk} is normally distributed for each aspect, with zero mean and variance s^2_{ehi}. The residual matrix at the score level is a diagonal matrix with (10 * 2) twenty elements.

At the second level, the mean of student *j* for every aspect is written as a deviation from the school mean:

$$ß_{h0jk} = g_{h00k} + u_{h0jk}. \qquad (2a)$$

In Equation (2a), g_{h00k} refers to the mean score for aspect *h* of school *k*. The random term, u_{h0jk}, indicates the deviation of student *j* of school *k* from this mean score. It is assumed that this residual score is uncorrelated with (e_{hijk}) of Equation (1) and that u_{h0jk} is normally distributed, with zero mean and variance s^2_{uhj}. The covariance matrix of the residual scores at the student level is a 10 * 10 symmetrical matrix with 55 elements (10 variances and 45 covariances).

At the third level, the mean of school *k* is written as a deviation from the population mean:

$$g_{h00k} = \pi_{h000} + v_{h00k}. \qquad (2b)$$

The fixed parameter π_{h000} is an estimate of the population mean. Because the scores are standardized (for each aspect and task), these (ten) fixed pa.ameters are of minor importance. It is assumed that the residual scores (v_{h00k}) are uncorrelated with the residual scores at both the student level (see: Equation 2a) and the residual scores at the score level (see: Equation 1). In addition, it is assumed that the residuals are normally distributed, with zero mean and variance s^2_{vhk}. The between-schools covariance matrix of the basic model is, just like the covariance matrix at the student level, a symmetrical matrix with 55 elements.

Simplified model (Model II)

Inspection of the covariance matrices of the basic model provided a first impression of the dimensionality of the data. Subsequently, an attempt was made to simplify the model without a loss of information. Two criteria were used. In order to maintain a factor as an independent entity within the model, its variance had to be significantly greater than zero ($p < .05$); if this was not the case, then the given factor was removed from that specific level of the model. Secondly, the correlation with all of the other factors had to be significantly smaller than one (or greater than -1); perfectly correlating factors were collapsed into a single factor. In the specific case at hand, the result was a drastically simplified model (Model II).

Task-effects model (Model III)

In Models I and II any possible differences between the tasks were disregarded at both the school and student level. In a preliminary analysis there appeared to be differences between the variances of the two tasks. In the task-effect model (Model III), these differences are taken into account. The influence of the writing task is analyzed simultaneously at the school and the student level. In order to do this, Equation (2b) is expanded at both levels using a dummy which represents the difference between both tasks. If $O2_{02jk}$ is a dummy-variable that is turned on for the letter of application, then:

$$ß_{h0jk} = g_{h00k} + u_{h0jk} + u_{00jk} * O2_{02jk}. \qquad (3a)$$

and,

$$g_{h00k} = \pi_{h000} + v_{h00k} + v_{000k} * O2_{02jk}. \qquad (3b)$$

Equations 1, 3a (or 2a) and 3b (or 2b) can also be represented as one formula. Substitution of Equation 3b in 3a, and the result in 1, gives:

$$Y_{hijk} = \Sigma_h \ (\pi_{h000} + u_{h0jk} + v_{h00k}) + O2_{02jk} \ (u_{00jk} + v_{000k}) + e_{hijk}. \qquad (4)$$

The task dummy ($O2_{02jk}$) has a value of 0 for the letter of apology and 1 for the letter of application. In the task-effect model (model III) the variances of the writing factors can then be interpreted as the variances of the letter of apology. The covariance matrix for a model with ten writing factors and one task factor (see Equation 4) at the student level can be represented as follows:

$$\begin{array}{lllll} s^2_{u1j} & & & & \\ s_{u1j,\ u2j} & s^2_{u2j} & & & \\ \cdot & & & & \\ \cdot & & & & \\ s_{u1j,\ u10j} & \cdots & \cdots & s^2_{u10j} & \\ s_{u1j,\ u0j} & \cdots & \cdots & s_{u10j,\ u0j} & s^2_{u0j} \end{array} \qquad (5)$$

In Equation (5), s^2_{u1j} is the variance of the first writing factor at the student level, $s_{u1j,\ u2j}$ represents the covariance between the first and second writing factor, s^2_{u2j} indicates the variance of the second writing factor, and s^2_{u10j} represents the variance of the tenth writing factor. The variance of the task factor is indicated by s^2_{u0j}; $s_{u1j,\ u0j}$ represents the covariance of this task factor with the first writing factor. The covariance matrix at the school level is written analogously to that at the student level (Equation 5). Note, however, that the number of factors at each level may be different. The variance of a factor for the letter of application is easily calculated as it is the sum of the variance of the writing factor, the covariance of this factor with the task factor multiplied by two, and the variance of the task factor. By way of example, the variance of the first writing factor of the letter of application is calculated as $(s^2_{u1j} + 2*s_{u1j,\ u0j} + s^2_{u0j})$.

In the task-effects model, the covariance between the writing factors represents the covariance for the letter of apology. The correlations between the writing factors can easily be calculated (by dividing the common covariance by the root of the product of the given variances). The correlations between the factors provide insight into the relation between the factors at the levels distinguished.

The ratio of the variances at the school and student levels provided insight into the distribution of the systematic variance between the two levels. As the scores were standardized for each task by aspect, the total variance for an individual factor is, in principle, equal to one.[2] The intraclass correlation, defined as the ratio of the between-schools variance in relation to the total variance, is then equal to the between-schools variance. Because the total variance contains a certain amount of error, the intraclass correlation is an underestimation of the 'true' intraclass correlation (Muthén, 1992). This 'error free' intraclass correlation can easily be calculated by dividing the between-schools variance by the total systematic variance (the sum of the between-schools variance and the between-students variance).

Reliability

The reliability of the factor means at the school level for factor *f* ($\rho_{f,\ school}$) can easily be calculated (Raudenbush, Rowan & Kang, 1991):

$$\rho_{f,\ school} = S^2_{vkf} \ / \ (S^2_{vkf} + S^2_{ujf} \ / \ N_j + S^2_{ehi} \ / \ N_i) \qquad (6)$$

for which

- S^2_{vkf}: the between-schools variance for factor *f*;

- N_j: the (mean) number of students per school (about 8);
- S^2_{ujf}: the between-students variance for this factor;
- N_i: the (mean) number of indicators;
- S^2_{ehi}: the residual variance (calculated as the average of the residual variances of the given indicators *h*).

The reliability of the factor means at the student level ($\rho_{f, student}$) can be calculated as:

$$\rho_{f,\ student} = S^2_{ujf} / (S^2_{ujf} + S^2_{ehi} / N_i) \qquad (7)$$

In the assessment report, writing performance was discussed for each task by aspect (see Kuhlemeier & Van den Bergh, 1989). The reliability of these scores can be calculated analogously by dividing the systematic variance ($S^2_{vkf} + S^2_{ujf}$) by the sum of the systematic variance ($S^2_{vkf} + S^2_{ujf}$) and the residual variance of the scores in question (S^2_{ehi}).

The multilevel analysis was conducted with the ML3-program (Prosser, Rasbash & Goldstein, 1987) using the IGLS-estimation method (Iterative Generalized Least Squares). The significance tests of the differences between variances and covariances (correlations) were conducted using a multiple comparison procedure (Goldstein, 1987, p. 29). If H0 is true, this procedure results in a statistic which is approximately Chi-square distributed. In statistical tests, the 5% significance criterion was maintained.

The relative fit of the task-effects model (Model III) as opposed to the analog model lacking this factor (Model II) was evaluated by the difference in -2 Log Likelihood between both models. This difference is asymptotically Chi-square distributed with the corresponding difference in the number of degrees of freedom (Bentler & Bonet, 1980). On the basis of this test, Model II was rejected in favour of Model III. It is with this latter model that the report of the findings begins[3].

Results

The number and the nature of the factors

How many factors have to be distinguished in order to give an adequate description of the writing achievements of secondary school students (first research question)? Table 1 shows the pattern matrix of the model which satisfactorily describes the data using the lowest posşible number of factors. The twenty test scores are plotted on the horizontal axis, the factors describing these scores are plotted on the vertical axis.

(insert Table 1 about here)

A total of eight factors are needed (see Table 1) to provide an adequate description of the relations between the twenty test scores. Two factors, FAC1 and TASK, are found at both the school and the student level so that the total number of different factors is six.

First, the structure at the school level has to be described. By temporarily disregarding the task factor, the original 10*10 matrix can be reduced to a 2*2 matrix. In the case at hand, only two writing factors are demonstrable. The first factor (FAC1) is indicated by the scores for global quality, content, style, organisation, and content elaboration. It is interpreted as the general-content quality of the writing products (defined in terms of characteristics of the writing products) or as composition skills (defined in terms of skills). The second factor (FAC2) is indicated by the five writing errors and is defined as the technical writing quality (technical conventions).

At the student level, the data can be described adequately using four writing factors and one task factor. The ten-aspect model postulated, with each aspect taking a separate factor, is not supported by the data. As at the school level, the first factor (FAC1) places demands on the general-content quality of writing and reflected the composition skills of the students. Unlike at the school level, however, the ten technical writing scores indicate three writing factors: punctuation (FAC3), spelling (FAC4) and a factor which is indicated by the scores for verb spelling and both grammatical measures (FAC5). This last factor is interpreted as the technical quality at the sentence level, as all three of the indicators are relevant to this level.

The factor means for composition and technical conventions vary systematically between schools and between students within schools. These differences are not independent of the writing task. At both the school and student level, the task factor contributes significantly to the description of the relationships between the twenty test scores. This means that the distinction of more than one writing task is supported by the data.

Differences between schools

At the school level, three factors emerged from the analyses which were reported on previously: composition, technical conventions, and a factor which indicated the difference between both letters. Important, in this regard, are the differences between schools (second research question). In order to deal with this issue, it is necessary to determine the between-schools variance of a given factor relative to the total variance (the intraclass correlation) or relative to the total systematic variance (the true or error-free intraclass correlation). The proportions of between-schools variance for the factors that were found at the school level (FAC1 and FAC2) are given in the upper section of Table 2.

(insert Table 2 about here)

In Dutch secondary education, there are relatively large differences between schools in the level of composition. Approximately one-fifth of the total variance in composition (21% for the letter of apology and 20% for the letter of application) can be attributed to the school the student is attending. The between-school differences in the factor means for technical conventions are markedly lower (5% and 10%, respectively). At the school level, the task only plays a role in technical writing: differences between schools on the letter of application are the highest for this factor (the difference between .05 and .10 is significant).

It should be noted that these statistics are an underestimation of the true differences between schools as measurement errors at the score level were not taken into account. After adjusting for unreliability, the proportions of between-schools variance are much higher. In the case of the letter of apology, for instance, 38% (composition) and 17% (technical conventions) of the systematic factor variation lies between schools.

It can be concluded that schools differ to a greater degree in the category of composition than they do in the category of technical conventions. The large differences between schools are not only attributable to differences in instructional effectiveness, but also reflect the highly selective character of Dutch secondary education in which students are sorted into different school types or tracks according to their achievements (Kreft, 1987).

Differences between students within schools

At the student level, five factors were found: composition, punctuation, spelling, a sentence factor indicated by spelling and the two grammatical standards, and a task factor for the difference between the two writing assignments. How great are the differences between students within schools (third research question)? The proportions of between-students variances for the four writing factors at the student level are given in the lower section of Table 2. It appears that the differences between students are greatest for composition (.35 for the letter of apology and .38 for the letter of application), followed by spelling (.30 and .33), punctuation (.23 and .25), and the sentence factor for verb spelling and grammar (.20 en .24).

At the student level, a task effect can be seen for three of the four writing skills factors, namely, composition, spelling, and the sentence factor; in all of the cases, the variance of the letter of application is the greatest.

Correlation between the factors

The correlations between the writing factors at the school and student level are also presented in Table 2 (fourth research question).

At the school level, the correlation between composition and technical conventions is -.79 for the letter of apology and -.43 for the letter of application. If a given school has a high factor mean for composition, it has a low factor mean for the factor that is indicated by the technical writing errors (and vice versa).

At the student level, an easy-to-interpret pattern can be seen. The correlations with the composition factor are all negative, while the correlations between the three technical convention factors are all positive. Skill in composition is almost totally independent of punctuation, spelling and technical quality at the sentence level. There are, however, two exceptions: the better a student composes a letter of apology, the fewer punctuation errors are made (r = -.16) and the fewer technical errors occur at the sentence level (r = -.15) he of she makes. For the letter of application there is no relation between the quality of composing and the technical quality.

It is striking that there is also little relation between the technical conventions. There is, however, one exception: the correlation between punctuation and spelling is relatively high (r = .73 for the letter of apology and r = .75 for the letter of application). Students who make few punctuation errors, likewise, do not often make spelling errors. This is completely consistent with current theories of writing (Flower & Hayes, 1980; Bereiter & Scardamalia, 1987).

Reliability at the school, student and score level

The residual variance, which cannot be accounted for by the factors at the school and student level, provides an initial indication of the (un)reliability with which the factors are measured (fifth research question). The proportions of the residual variance by aspect per task are shown in Table 3.

(insert Table 3 about here)

A large portion of the observed variance is attributed to the score level (see Table 3). Stated otherwise: relatively few variance is attracted by the factors at the school and student level.

Not all of the twenty test scores are equally useful for indicating the factor they are linked to. Depending on the aspect and the task, 23% to 93% of the total variance is due to differences in students' scores. There is, for example, a striking heterogeneity in the composition component. The global quality and organisation of the letters are the best indicators of the composition factor; the rating of the content and style are the worst indicators. The residual variances of the technical writing scores are all high (50% or higher). The measurement of verb spelling consists primarily of error or random noise (88% and 93% for the letter of apology and the letter of application, respectively). It would appear that technical conventions are more difficult to measure than composition skills (for a similar result, see Breland et al., 1987).

A multiple comparison procedure (Goldstein, 1987, p. 29) was used to determine whether the residual variances differ from task to task. The hypothesis that the variances of both tasks are equal has to be rejected (p < .05). Inspection of the individual contrasts revealed that the aspect style is a much better indicator of composition in the letter of apology than it is in the letter of application. The reverse is true of the number of content elements and the proportional number of spelling errors. There, the residual variance is greater in the letter of apology. In short, the degree to which the test scores indicate the writing factors is not invariant across tasks. Estimation of two residual variances for each aspect, one for the letter of apology and one for the letter of application, proves to be useful.

Information on the reliability of the measurement at the school, student, and score level (see Equation 6 and 7) is shown in Table 4 (fifth research question).

(insert Table 4 about here)

If the criterion for reliable measurement is, rather arbitrarily, set at .70, the reliability of the school factor means for the composition factor is not sufficient (.63 for the letter of apology and .60 for the letter of application). The reliability of the school factor means for technical conventions is quite low (.21 and .35, respectively).

At the student level, only the composition factor is measured with a sufficient degree of reliability (.81 and .82). The reliability of the measurement of punctuation, spelling, and the sentence factor is rather low.

The reliability of the test scores, the report units used in the Dutch assessment of language performance, is often extremely low. The range is from .22 to .71, with an average of .45.

Discussion

In recent language assessments, writing performance is measured and reported using a large number of text characteristics in order to evaluate the performance on various writing assignments. In this study, the efficiency of this comprehensive and differentiated measurement and reporting is assessed. It appears that the relationships between the twenty test scores, in this case reporting-units, can be described adequately using only five writing factors and one task factor.

The number and the nature of the factors is not equal at each level. At the school level, three factors are needed in order to give an adequate description of the relationships between the twenty test scores: two writing factors and one task factor. The first factor concerns the general-content quality of the writing, in other words, the quality of composition; the second involves the technical quality of the writing. Approximately one-fifth of the total variance in composition (21% for the letter of apology and 20% for the letter of application) can be attributed to the school the student is attending. The between-school differences in the factor means for technical conventions are markedly lower (5% and 10%, respectively). These factor variances at the school level can be attributed to differences between schools in, among other things, intake characteristics, curricular offerings, and the effectiveness of instruction and school policy (Longford & Muthén, 1990; Melse & Kuhlemeier, 1993).

The dimensionality at the student level is not identical to that at the school level. At the student level, five factors are necessary in order to give an adequate description of the data: four writing factors and one task factor. Just as at the school level, a composition factor can be found; the student does not add any new dimensions. In contrast to the findings at the school level, the technical conventions factor is divided into three subskill factors: punctuation, spelling, and a factor which is interpreted as technical quality at the sentence level. Consistent with the statistical modelling, the structure at the student level can be regarded as a deviation from the common structure at the school level. The structure at the student level represents the extent to which individual students deviate from this common structure. The proportion of between-student factor variances varies with the factor and the task from .20 tot .38. These student level factor variances refer to the unique, idiosyncratic contribution of the student, due to, for instance, his or her intelligence, perseverance, and motivation. With the exception of the relationship between punctuation and spelling, the correlations between the student-level factors are rather low, the quality of composition is almost totally independent of the technical quality.

The reporting of writing performance

The structure of writing performance has proved to be task-dependent at the school, student, and score level. The variance of the factors and the correlations between them differ from task to task, while the proportions of the residual score variance also differ. Thus, the administration and reporting of more th... one writing task is supported by the data.

The measurement and reporting of ten writing aspects is not supported by the data. If the goal is to describe the mean writing skills of the students in the population, only the three school level factors would be important (composition, technical conventions and a task factor). If one is also interested in differences between students, then five factors have to be distinguished (e.g., factors for composition, punctuation, spelling, the technical quality at the sentence level, and a task factor). From a purely psychometric perspective, the reporting of the writing achievements in the Dutch Assessment of Language Performance (Kuhlemeier et al., 1989) is unnecessary complicated.

Measurement of writing

In the language assessments, writing performance is generally measured using specific assignments (Schoonen, 1991). Unlike the free assignment, this specific assignment has a number of prescribed elements which involve certain aspects of the communicative context such as the goal and the audience. The content, style, organisation, and the more technical aspects of the text are generally determined by the student. Like other performance-based tests, the face validity appears to be high (Burger & Burger, 1994). The efficiency in measuring writing performance is, however, questionable.

Firstly, the five indicators for composition skills prove to be indistinguishable. The scores for general quality, content, style, organisation, and content elaboration form a single factor. The content, stylistic, and organisational aspects would probably be measured more efficiently by tasks that are structured to focus on those aspects. That is not to say that these tasks are by definition less valid than free or specific assignments (Breland et al. 1987; Van Schooten & De Glopper, 1990; Schoonen, 1991). Nonetheless, it is not argued in this article that specific assignments should be ruled out in national assessment. The face validity of this test form is too high to disregard it completely in national assessment (Schoonen, 1991). What can be eliminated in the future is the costly and labour-intensive rating of the multiple composition aspects. It is sufficient to use the most reliable indicator of composition skills which also requires the least amount of training and assessment time, the assessment of global quality, operationalized as the general impression gained by the reader after a single reading.

Secondly, it has become clear that the specific assignment is not an efficient instrument for measuring the technical aspects of writing. The punctuation, spelling, and sentence factors proved to be distinguishable. The reliability, however, was so low that the use of specific assignments for this purpose is not advisable. Subskill tests, such as multiple-choice assignments, correction assignments, and dictation would be more suitable (Culpepper & Ramsdell, 1982; Van Schooten & De Glopper, 1990).

National assessment and school effectiveness

In the study of school effectiveness, the ability to establish differences between schools is a minimal requirement. In this study, secondary schools differed substantially in levels of composition skills while differences in the technical quality of the writing were much smaller. The reliability of the measurement at the school level proved to be low. If the criterion for adequate measurement is set at .70, then the school factor means for composition and technical quality were measured with a insufficient degree of reliability to justify secondary analysis or follow-up studies. There are a number of possible reasons for this. One is the low number of students per school. The most efficient method of gaining a precise estimate of the population indices is a data-collection design with one student per item per school. In national assessment the total number of students sampled within a school is usually small. In addition, subsets of items are administered to small subsamples of the students within a school (Beaton, 1988; Rust & Johnson, 1992) in matrix sampling designs (Sirotnik, 1974). Given the primary goal of national assessment, the precise estimation of population indices, this is a reasonable choice. A major drawback to this method is, however, that it becomes difficult (Kuhlemeier, 1994) if not impossible (Rijlaarsdam, Van den Bergh & Zwarts, 1992) to distinguish school and student variances. The power to estimate the between-school variances is low (Bosker & Snijders, 1990), and thus the standard error of the estimate will be large relative to the estimate. If school effectiveness research is to be linked to national assessment (Scheerens & Stoel, 1987; Bock, 1988) it is recommended that the existing sample and administration designs be reviewed.

Data analysis

The last recommendation concerns the statistical analysis. National assessment data are generally analyzed using unilevel techniques. If adjustments are made for the design effect, this is done using jack-knife and weighting procedures (Johnson & Rust, 1992). On the basis of the findings in this specific investigation, it would be wise to model the hierarchical structure of the data explicitly in a multilevel analysis.

Notes

1 The scores on the different tasks are seen in this connection as fixed effects. As was shown by Cla.k (1979), it can be argued that these effects should be seen as random effects. In that case, the tasks would form a random sample of the population of possible tasks to which they could be generalized. This would imply that no task-per-aspect averages would be reported in the national assessment reports, instead aspect averages with a range for the spread resulting from different tasks would be used.

2 The variances at the school, student, and score level reported in the tables do not always add up to 1.00. This is partially due to the rounding off of numbers and also to the manner in which the factor variances are modelled. The latter requires some explanation. The variance of a factor at a

given level is a weighted average of the variance of the indicators at the same level. The between-schools variances of global quality, style, organisation, and content elaboration, for example, determine the variance of the common composition factor. The variances of the indicators are not, however, all equal. Th between-schools variance for global quality, content, and style, for example, is significantly higher than the variance for content and content elaboration. In the analysis, such differences are accounted for by modelling a separate variance in addition to the factor. In the example discussed here, a separate variance for content elaboration was estimated in addition to the composition skills factor. This led to a slight change of the factor variance which in turn caused the total variance to be different from 1.00.

3 An important assumption is the (multivariate) normality of the random terms at each level. The distribution of the technical error scores is rather skewed (see Appendix), because a relatively small group of students is responsible for almost all of the errors in punctuation, spelling and grammar, while the vast majority of Dutch students make no or only a few technical errors. To check the normality assumption, the test scores were separately normalized using the normalizing transformation described in Prosser, Rasbash and Goldstein (1991, p. 44) and the data were reanalysed. In general, the effects on the variances and covariances appeared to be minimal. The pattern of outcomes remained the same. For this reason, the original data are used in this article.

4 As the scores are standardized by task per aspect, with a mean of zero and a variance of one, the systematic variance of a factor is approximately equal to the total variance minus the (weighted) average at the residual variances of the scores which indicate the given factor.

References

Ackerman, T.A. & Smith, P.L. (1988). A comparison of information provided by essay, multiple choice, and free-response writing tests. *Applied Psychological Measurement, 12*, 117-128.

Applebee, A.N., Langer, J.A., Mullis, I.V.S. & Jenkins, L.B. (1990). *The writing report card, 1984-1988*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Balke, G. (1991). *Multilevel factor analysis of proficiency in English as a foreign language*. Paper presented at the symposium 'Multilevel factor analysis: applications to education' at the annual meeting of the American Educational Research Association, Chicago, April 3-7, 1991.

Beaton, A. (1988). *Expanding the new design: the NAEP 1985-1986 technical report*. Princeton, NJ: Educational Testing Service.

Bentler, P.M. & Bonet, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bergh, H. van den (1988). *Examens geëxamineerd* [National exams examined]. 's-Gravenhage: SVO.

Blok, H., & Eiting, M.H. (1988). De grootte van schooleffecten: hoe verschillend presteren leerlingen van verschillende scholen? [The size of school effects: how different is the achievement of students from different schools?]. *Tijdschrift voor Onderwijsresearch, 13*, 16-30.

Bock, R.D., Mislevy, R.J. & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11*, 3, 4-11.

Bock, R.D. (1988). A design for a biennial national education assessment, reporting by state. *International Journal of Educational Research, 12*, 745-750.

Bosker, R.J. & Snijders, T.A.B. (1990). Statistische aspecten van multiniveauonderzoek [Statistical effects of multilevel research]. *Tijdschrift voor Onderwijsresearch, 15*, 317-329.

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M. & Rock, D.A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examination Board.

Burger, S.E. & Burger, D.L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice, 13*, 1, 9-15.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Conry, E. & Jerolski, S. (1980). *The British Colombia assessment of written expression. General report*. Province of British Colombia: Ministry of Education.

Culpepper, M.M. & Ramsdell, R. (1982). A comparison of a multiple choice and an essay test of writing skills. *Research in the Teaching of English, 16*, 295-297.

Flower, L.S. & Hayes, J.R. (1980). The dynamics of composing: making plans and juggling with constraints. In: L.W. Gregg & E.R. Steinberg (Eds.). *Cognitive processes in writing* (pp. 31-50). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cipps, G. & Goldstein, H. (1983). *Monitoring children. An evaluation of the Assessment of Performance Unit*. Educational Books: London.

Glopper, K. de (1988). *Schrijven beschreven. Inhoud, opbrengsten en achtergron den van het schrijfonderwijs in de eerste vier leerjaren van het voortgezet onderwijs* [Writing described. Content, results and backgrounds of writing education in the first four years of secondary education]. 's-Gravenhage: SVO.

Goldstein, H. (1987). *Multilevel models in educational and social research.* Charles Griffin: London.

Goldstein, H. & McDonald, R.P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 4, 455-467.

Gorman, T.P., White, J., Orchard, L., Tate, A. & Sexton, B. (1981). *Language performance in schools; Primary survey report no. 1.* London: Her Majesty's Stationary Office.

Hox, J.J. (1993). Factor analysis of multilevel data: gauging the Muthén model. In: J.H.L. Oud & R.A.W. van Blokland-Vogelesang (Eds.). *Advances in longitudinal and multivariate analysis in the behavioral sciences* (pp. 141-155). Nijmegen: ITS.

Johnson, E.G. & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175-190.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-132.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Kreft, I. (1976). *Models and methods for the measurement of school effects.* Amsterdam: University of Amsterdam.

Kuhlemeier, J.B. (1994). Effecten van schoolkenmerken en steldidactiek op functionele stelvaardigheid Nederlands in het MAVO [Effects of school characteristics and writing didactics on Dutch functional writing performance in intermediate general education]. *Tijdschrift voor Onderwijsresearch, 19*, 2, 161-181.

Kuhlemeier, J.B. & Bergh, H. van den (1989). *De proefpeiling Nederlands: een onderzoek naar de haalbaarheid van peilingsonderzoek in het voortgezet onderwijs* [National assessment of language performance: a feasibility study in secondary education]. Arnhem: National Institute of Educational Measurement.

Kuhlemeier H. & Bergh, H. van den (1991). De correlationele structuur van taalvaardigheid: een exploratie [The correlational structure of language abilities: an exploration]. *Tijdschrift voor Onderwijsresearch, 16*, 143-159.

Kuhlemeier, J.B., Bergh, H. van den, Notté, H., Wagenaar, H., Verstralen, H. & Cappers, R. (1994). *Balans van het aardrijkskunde-onderwijs in het derde leerjaar van het voortgezet onderwijs* [State of the art of geography education in the ninth grade of secondary education]. Arnhem: National Institute of Educational Measurement.

Linn, R.L. & Burton, E. (1994). Performance-based assessment: implications of task specificity. *Educational Measurement: Issues and Practice, 13*, 1, 5-8.

Longford, N.T. (1990). Multivariate variance component analysis: an application in test development. *Journal of Educational Statistics, 1990, 15*, 2, 91-112.

Longord, N.T. & Muthén, B.O. (1992). Factor analysis for clustered observations. *Psychometrika, 57*, 4, 581-597.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley Publishing Company, Inc.

Melse, L. & Kuhlemeier, J.B. (1993). De structuur van schrijfvaardigheid Duits: een toepassing van multilevel factoranalyse [The structure of writing ability in German as a second language: an application of multilevel factor analysis]. In: C. Blankenstijn & A. Scheper (Eds). *Taalvaardigheid: Symposiumbundel werkverband Amsterdamse psycholinguisten. Wap Publikatie 2* [Language Ability: Congress volume Amsterdam psycholinguistics workgroup] (pp. 139-152). Dordrecht: ICC Publications.

Muthén, B.O. (1992). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354

Neville, M. (1988) *Assessing and teaching language: Literacy and oracy in schools*. London: Macmillan.

Prosser, R., Rasbash, J. & Goldstein, H. (1991). *ML3: Software for three-level analysis. User's guide*. London: Institute of Education.

Quellmalz, E.S., Capell, F.J. & Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*, 241-258.

Raudenbush, S.W., Rowan, B. & Kang, S.J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16*, 4, 295-330.

Robinson, S.W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351-357.

Rust, K.F. & Johnson, E.G. (1992). Sampling and weighting in the national assessment. *Journal of Educational Statistics, 17*, 111-129.

Rymenans, R., Leroy, G. & Daems, F. (1991). *Geletterdheid op achttien jaar: peiling naar de lees- en schrijfvaardigheid bij het einde van het secundair onderwijs. Deel II: Functionele taalvaardigheid* [Language performance at eighteen: an assessment of reading and writing skills at the end of secondary education. Part II: Functional language skills]. Antwerpen: Universitaire Instelling Antwerpen, Departement Didactiek en Kritiek.

Rijlaarsdam, G., Bergh, H. van den & Zwarts, M. (1992). Incidentele transfer bij produktieve taalopdrachten: een aanzet tot een baseline [Incidental transfer in productive language assignments: a baseline study]. *Tijdschrift voor Onderwijsresearch, 17*, 55-66.

Scheerens, J. & Stoel, W.G.R. (1987). Conceptuele en methodologische problemen bij onderzoek naar de effectiviteit van schoolorganisaties [Conceptual and methodological problems in research on the effectiveness of school organisations]. *Bijdragen aan de Onderwijsresearch [Contributions to educational research]*, 1-16.

Schoonen, R. (1991). *De evaluatie van schrijfvaardigheidsmetingen* [The evaluation of the measurement of writing]. Amsterdam: Universiteit van Amsterdam.

Schooten, E. van & Glopper, K. de (1990). De validiteit van meerkeuze-instrumenten voor het meten van schrijfvaardigheid [The validity of multiple choice instruments for the measurement of writing ability]. *Tijdschrift voor Taalbeheersing, 12*, 93-110.

Sirotnik, K. (1974). An introduction to matrix sampling for the practitioner. In: W.J. Popham (Ed.), *Evaluation in education: current applications* (pp. 453-529). Berkeley: McCutchen.

Sytstra, J. (1993). *Balans van het taalonderwijs halverwege de basisschool*. [State of the art of language education in the first half of primary education]. Arnhem: National Institute of Educational Measurement.

Sytstra, J. & Zwarts, M.A. (1991). *Constructvaliditeit van het model van domeinbeschrijvingen. Verslag van een secundaire analyse* [Construct validity of the model of content description. Report on a secondary analysis]. Arnhem: National Institute of Educational Measurement.

Tate, R.L. & King, F.J. (1994). Factors which influence precision of school-level IRT ability estimates. *Journal of Educational Measurement, 31*, 1, 1-15.

Vergeer, M.M. & Oostdam, R.J. (1988). *Voorstudie periodieke peiling van het onderwijsniveau in het speciaal onderwijs. Deel 3: prestaties van LOM- en MLK-leerlingen op functionele taaltaken* [Feasibility study on assessment in special education. Part 3: achievement of LOM and MLK students on functional language assignments]. Amsterdam: SCO.

Wesdorp, H., Bergh, H. van den, Bos, D.J., Hoeksma, J.B., Oostdam, R.J., Scheerens, J. & Triesscheijn, B. (1986). *De haalbaarheid van* *peilingsonderzoek; een voorstudie op het gebied van het taalonderwijs in de lagere school* [The feasibility of national assessment; a preliminary study in the realm of language education in primary schools]. Lisse: Swets & Zeitlinger.

Wijnstra, J.M. (1988). *Balans van het rekenonderwijs in de basisschool* [State of the art of mathematics eduation in primary education]. Arnhem: National Institute of Educational Measurement.

Zwarts, M. (1990). *Balans van het taalonderwijs aan het einde van de basisschool* [State of the art of language education at the end of primary education]. Arnhem: National Institute of Educational Measurement.

Zwarts, M. & Rijlaarsdam, G. (1991). *Verantwoording van de taalpeiling einde basisonderwijs 1988* [Scientific report on language assessment at the end of primary education]. Arnhem: National Institute of Educational Measurement.

Table 1 *Pattern matrix at the school and student level (model III; APOL: letter of apology; APPL: letter of application)*

	Between-schools			Between-students				
	FAC1	FAC2	TASK	FAC1	FAC3	FAC4	FAC5	TASK
Composition								
1.1 Global quality: APOL	1			1				
1.2 Global quality: APPL	1		1	1				1
2.1 Content: APOL	1			1				
2.2 Content: APPL	1		1	1				1
3.1 Style: APOL	1			1				
3.2 Style: APPL	1		1	1				1
4.1 Organisation: APOL	1			1				
4.2 Organisation: APPL	1		1	1				1
5.1 Content elaboration: APOL	1			1				
5.2 Content elaboration: APPL	1		1	1				1
Technical conventions								
6.1 Punctuation: APOL		1			1			
6.2 Punctuation: APPL		1	1		1			1
7.1 Spelling: APOL		1				1		
7.2 Spelling: APPL		1	1			1		1
8.1 Spelling of verbs: APOL		1					1	
8.2 Spelling of verbs: APPL		1	1				1	1
9.1 Grammar 1: APOL		1					1	
9.2 Grammar 1: APPL		1	1				1	1
10.1 Grammar 2: APOL		1					1	
10.2 Grammar 2: APPL		1	1				1	1

Table 2 *Factor variances (on the diagonal) and correlations (below the diagonal) for the writing factors at the school and student level for the letter of apology and the letter of application*

WRITING FACTORS		LETTER OF APOLOGY				LETTER OF APLLICATION			
					SCHOOL LEVEL				
		FAC1	FAC2			FAC1	FAC2		
FAC1	Composition	.21				.20			
FAC2	Techn. conventions	-.79	.05[1]			-.43	.10[1]		
					STUDENT LEVEL				
		FAC1	FAC3	FAC4	FAC5	FAC1	FAC3	FAC4	FAC5
FAC1	Composition	.35[1]				.38[1]			
FAC3	Punctuation	-.16	.23			-.06[2]	.25		
FAC4	Spelling	-.08[2]	.73	.30[1]		-.03[2]	.75	.33[1]	
FAC5	Verb spelling/gramm.	-.15	.43	.36	.20[1]	-.00[2]	.50	.45	.24[1]

[1] The variance of the letter of apology differs significantly from that of the letter of appliaction;

[2] the covariance or correlation is not significantly different from zero.

Table 3 *Residual variances by aspect per task*

	Letter of apology	Letter of application
Composition		
1 Global quality	.23	.25
2 Content	.57	.64
3 Style	.39	.61[1]
4 Organisation	.24	.26
5 Content elaboration	.62	.41[1]
Technical conventions		
6 Punctuation	.67	.66
7 Spelling	.64	.50[1]
8 Spelling of verbs	.88	.93
9 Grammar 1	.71	.67
10 Grammar 2	.76	.74

[1] The variance of the letter of application differs significantly from that of the letter of apology.

Table 4 *Reliability of the measurement at school-, student- and score level*

		Letter of apology	Letter of application
School level			
FAC1	Composition	.63	.60
FAC2	Technical conventions	.21	.35
Student level			
FAC1	Composition	.81	.82
FAC3	Punctuation	.26	.27
FAC4	Spelling	.32	.40
FAC5	Verb spelling/grammar	.43	.48
Score level			
1	Global quality	.71	.70
2	Content	.50	.48
3	Style	.59	.49
4	Organisation	.70	.69
5	Content elaboration	.47	.59
6	Punctuation	.29	.35
7	Spelling	.35	.46
8	Verb spelling	.22	.27
9	Grammar 1	.26	.34
10	Grammar 2	.25	.31

Appendix

Sample means, standard deviations, skewness, range and number of students by aspect per task (APOL: letter of apology; APPL: letter of application)

Variable	Mean	Std Dev	Skewness	Minimum	Maximum	N
Composition						
Global quality: APOL	8.90	2.89	-.13	3	15	1436
Global quality: APPL	7.46	2.79	.36	3	15	1430
Content: APOL	13.10	2.38	-1.64	3	15	1429
Content: APPL	9.55	4.44	-.41	3	15	1434
Style: APOL	7.88	2.60	.17	3	15	1422
Style: APPL	5.94	2.49	.82	3	15	1434
Organisation: APOL	9.09	2.91	-.09	3	15	1429
Organisation: APPL	8.34	2.98	.17	3	15	1438
Content elaboration: APOL	17.17	3.89	-.22	1	37	1425
Content elaboration: APPL	25.63	10.28	.19	1	53	1437
Technical conventions						
Punctuation[1]: APOL	82.24	64.30	1.55	0	450	1434
Punctuation[1]: APPL	100.10	77.40	2.00	0	800	1436
Spelling[2]: APOL	2.81	3.10	1.99	0	21	1432
Spelling[2]: APPL	4.76	4.12	1.66	0	31	1434
Spelling[3] of verbs: APOL	3.55	9.71	3.71	0	100	1433
Spelling[3] of verbs: APPL	2.49	7.70	4.87	0	100	1427
Grammar 1[4]: APOL	42.79	40.18	1.65	0	350	1436
Grammar 1[4]: APPL	44.71	38.64	1.80	0	400	1425
Grammar 2[5]: APOL	63.83	30.86	-.21	0	100	1435
Grammar 2[5]: APPL	61.03	28.66	-.21	0	100	1429

[1] number of punctuation errors per hundred sentences;

[2] number of spelling errors per hundred words;

[3] number of verb-spelling errors per hundred sentences;

[4] number of grammatical and idiomatic errors per hundred sentences;

[5] number of grammatically correct sentences per hundred sentences (in the multilevel analysis the test score is multiplied by minus one).