

R. Schoonen (1991). *De evaluatie van schrijfvaardigheidsmetingen. Een empirische studie naar betrouwbaarheid, validiteit en bruikbaarheid van schrijfvaardigheidsmetingen in de achtste groep van het basisonderwijs*. Amsterdam: Universiteit van Amsterdam/SCO.

The Evaluation Of Writing Measurements. An empirical study of the validity, reliability and utility of writing measurements in primary education (12 yrs).

SUMMARY

Introduction

One of the aims of the National Assessment of Educational Performance (Dutch: Periodieke Peiling van het Onderwijsniveau) in The Netherlands is to assess the writing ability of the students in primary education. In a feasibility study preparatory to the National Assessment, Wesdorp and his colleagues (1986) stated that the assessment of writing ability by means of writing tasks (direct assessment) needed further investigation. For example, it was difficult and costly to rate the students' texts reliably. Using more structured writing tasks and rating procedures could help avoiding some of the assessment problems. However, structured tasks and rating procedures threaten the validity of the writing assessment. This balance between reliability and utility on the one hand and validity on the other hand is the main focus of this study.

In this study five different types of writing task were evaluated. The writing tasks represented different degrees of structuring. Two tasks could be regarded as direct measurements, two as semi-direct measurements and one as an indirect measurement. With the most free task, the students receive a full specification of the rhetorical situation of the writing assignment (further referred to as the 'specified task'). In the second direct measurement, the rhetorical specification is supplemented with hints supporting the students' representation of the task and the generation and organization of the required text (this task is further referred to as the 'structured task'). In two semi-direct measurements students are provided with texts that supposedly fulfill the requirements of the rhetorical situation, but need to be revised. In one of these two tasks the erroneous parts of the text are underlined to draw the students' attention to them (these semi-direct tasks are further referred to as the 'interlinear revision task' and the 'interlinear revision task with marked errors'). The indirect measurement is a 'multiple choice task'.

According to the Hayes & Flower model (1980), the writing tasks should (cumulatively) suppress several writing processes. That is, the specified task requires students to generate and organize relevant information for their text, whereas the structured task suppresses these processes by offering support (i.e. hints) in the task. In the multiple choice task the student only has to choose between a few optional formulations. Obviously, the smaller the number of writing processes involved in the task, the fewer the rating problems. The multiple choice task creates almost no rating problems and is economic, whereas the rating of the specified task is subject to several different problems and is costly. The structured and interlinear tasks fall between these two extremes.

Another way to improve rating reliability is to structure the rating procedure as opposed to structuring the writing task. Two scoring methods were compared with respect to the ratings they produced for two tasks: the specified and the structured task. The methods

consisted of (1) essay scales with example essays as 'range finders', and (2) scoring according to strict scoring guides (counts). The structuring of both the writing task and the rating procedure were expected to simplify the assessment of writing ability. This possible simplification introduces an additional factor: the experience and professionalism of the rater. Simpler assessment techniques were expected to make it possible for less professional, and thus less expensive, raters to do the work equally reliably.

The present study investigated the question whether these different kinds of writing tasks and rating procedures may be considered a reliable and valid operationalization of writing ability (cf. chapter 1). Previous writing research is not conclusive about the effects of different tasks, rating procedures and levels of expertise of raters (cf. chapter 2). With respect to the writing tasks, Quellmalz et al. (1982) and Ackerman & Smith (1988) have each proposed a model to account for the interrelations between the different tasks. These models and the model of decreasing complexity (i.e. simplex model) and a model assuming the same construct representation for all assignments, were fitted to the data as a way to establish the construct validity of the writing assignments. The construct validity of the assignments was also evaluated by relating the different writing scores to intellectual ability tests. These tests were developed within the framework of Guilford's 'Structure-of-Intellect'(SI) model.

The criterion oriented validity was also evaluated in two different ways. The criterion validity was assessed by relating the different writing performances to the judgement of the students' teachers and by estimating the instrument bias, i.e. the interaction between the type of writing assignment and two group characteristics: sex and writing ability. The writing ability was measured with the specified task rated according to essay scales, as was done in Wesdorp's feasibility study.

Methods

Specified tasks and some of the rating instruments were already available from Wesdorp's feasibility study. The instruments previously not available were developed and pretested in the present study. In order to make unbiased comparisons between the writing tasks, all tasks concerning texts (i.e. the specified, the structured and the two kinds of interlinear tasks) were developed with the same rhetorical specification (i.e. topic, goal, audience etc. were identical). These sets of writing tasks were developed for four different rhetorical specifications (cf. chapter 3). The multiple choice test was not related to a rhetorical specification.

Twenty-two schools with a total of 442 students (ca. 12 years old; Dutch: 'groep 8') participated in the validation study. The students (within one class) were randomly assigned to five conditions. Four of the five groups in a class completed all four writing tasks for one of the four rhetorical specifications. The fifth group was a control group which completed only the specified tasks but for all four rhetorical specifications. All groups completed the multiple choice test. A large subsample of the students also completed the SI ability tests.

Thirty raters rated students' writings and revisions. They worked (independently) in panels of five (for the specified and structured task) or in panels of three (for both kinds of interlinear tasks). Two aspects of the students' essays were rated: (1) content and organization and (2) language usage. Both interlinear tasks were constructed so as to contain 'errors' in these two areas, and the multiple choice test also contained items relating to both aspects.

Data collection and rating procedures were shown to be satisfactory. Rating reliability was high. The multiple choice writing tests indicated an adequate level of internal consistency. However, most of the intellectual ability tests were less reliable. This was considered an handicap for the construct validation of the writing tasks (cf. chapter 5).

Results

The effect of rater expertise was evaluated by comparing the rater reliability of a panel of experts with that of a panel of non-experts. The rating procedures were 'essay scales' and 'scoring guides'. The panels also judged the text revisions collected using the two different kinds of interlinear tasks. Although in none of the points of comparison could the two types of panels be considered as (psychometrically) parallel, the non-expert panel did reach satisfactory rater reliability. In some cases the rater reliability of the non-expert panel was even slightly higher than that of the expert panel. However, the experts were evidently more reliable raters than the non-experts in one significant aspect: the rating on the scales of language usage in texts (cf. chapter 4).

The effects of writing task and rating procedure on rating reliability were investigated using the ratings of expert panels in the validation study. The results did not completely support the theoretical assumptions. The texts of the structured tasks were rated in fact no more reliably than the texts of the specified tasks. Ratings of the revisions in the interlinear task 'with marked errors' were even less reliable than those of the unmarked version. However, as expected, the interlinear tasks were rated far more reliably than the specified tasks and the structured tasks. Furthermore, the rating 'with scoring guides' was shown to be more reliable than the rating with essay scales. This difference between the two rating procedures was related to the aspect of the texts being rated; rating reliability appeared to vary between the rating procedures when content and organization was being rated, but there was no such difference when rating language usage.

The writing assignments showed different criterion validity, using the teachers' judgement as the criterion. Scores for the specified task showed higher correlations with the criterion than the scores for the structured task. The scores for the interlinear task 'without marking' also showed higher correlations with the criterion than the interlinear task with marked errors. These findings were in accordance with our expectations. However, a comparison of the criterion validity across the specified task, the interlinear task 'without marking' and the multiple choice test did not produce the expected result: the multiple choice test scores and the scores for the interlinear task showed a criterion validity as high as the validity of the scores for the specified task. The rating procedures correlated differently with the teachers' judgement: the scale ratings correlated more highly with the teachers' judgement than the ratings with the scoring guides and therefore appear to be more valid.

The analysis of instrument bias further supported this differential validity of the writing assignments. Valid writing assignments should not show any bias with respect to sex and writing ability. No bias with respect to sex could be demonstrated, but there was bias with respect to writing ability. Proficient writers seemed to be unable to produce a distinctive performance in the interlinear and multiple choice tasks, whereas less proficient writers seemed especially to benefit from the 'simplicity' of these more restrictive tasks. These findings appeared most strongly in relation to the content and organization scores.

Further investigation of the construct validity of the writing assignments showed that the assignments could not be described in a model that assumes the same factorial structure for all the different types of assignment. The simplex model that assumes the decreasing complexity of the writing assignments going from the specified task to the multiple choice task did not fit the data either. Two recently developed models (of Ackerman & Smith (1988) and Quellmalz et al. (1982)) seemed to fit the data reasonably well. The Aspect-simplex model of Ackerman and Smith turned out to be the most plausible of the two models. This model assumes that the different assignments measure the same aspects (content and organization and language usage) and that they also indicate three steps of decreasing complexity, the specified task being the most complex, the multiple choice *and* both the interlinear tasks being the least complex and the structured task falling in between. This range of complexity was demonstrated most clearly with the content and organization scores. Language usage scores demonstrated a less clear structure.

These last findings were supported by the correlations of the writing scores with the intellectual ability tests. The writing assignments correlated differently with the ability tests. The content and organization scores of the specified task showed substantial correlations with a broad range of ability tests, whereas the multiple choice test and the interlinear task correlated with a smaller range of ability tests. The different language usage scores did not show such clear patterns of correlation with the ability tests (cf. chapter 6).

Conclusion

Although this study did not consider all parameters relevant to assessment studies (such as backwash effects etc.), one must conclude that multiple choice and interlinear tasks are inadequate assignments to assess writing ability with a high level of validity. Content and organizational aspects of texts seem to be best assessed by specified tasks. Language usage can also be assessed using the more restrictive tasks such as the multiple choice and the interlinear task. The rating of the essays is best done using essay scales rather than strict scoring guides (counts). Non-expert raters appeared to be able to use the essay scales reliably as long as they had to rate content and organization; in rating language usage the non-experts were less reliable than the experts (chapter 7).