The assessment of writing ability: expert readers versus lay readers

Rob Schoonen, Margaretha Vergeer and Mindert Eiting* SCO-Kohnstamm Institute for Educational Research, University of Amsterdam

This article reports on three studies about the reading reliability of lay and expert readers in rating three kinds of writing assignments. Readers had to rate the Content and Language Usage of students’ writing performances. The studies show that expert readers are more reliable in rating Usage, whereas both lay readers and expert readers prove to be reliable raters of Content. When the writing assignment becomes more restrictive, as is the case with interlinear revision tasks, lay readers are also reliable raters of Usage. The conclusion is that the effect of expertise is dependent on the rating task the readers have to perform. Expertise pertains most to language usage and to the cases in which the writing assignments as well as the scoring instructions are relatively free.

I Introduction

The assessment of students’ writing ability is often an important part of (national) assessments (cf. APU – Assessment of Performance Unit (UK) and NAEP – National Assessment of Educational Progress (USA)). The direct measurement of writing ability is in those assessments considered the most valid way to gather information about the general level of the students’ writing proficiency. This means that students have to perform ‘real’ writing tasks. Subsequently the students’ writing performances have to be rated by readers. Writing assessments are therefore expensive and time-consuming. But whereas writing assessment through tests is much disputed in education, writing assessment through the rating of essays is generally accepted as a valid method (Cooper and Odell, 1977; Greenberg, Wiener and Donovan, 1986). Even so, this method of assessing writing ability is not without its problems. In the past three decades, numerous studies have been conducted to estimate score reliability and reading reliability as functions of the number of essays which should be assigned to a writer and the number of readers who should read essays (Godshalk, Swineford and Coffman, 1966; Wesdorp, 1974). Scores for different essays often show low correlations and readers are not always consistent in their judgements nor do they
always agree with other readers. Thus, the assumed validity of direct writing assignments is threatened by the unreliability of the scores, caused by the task variability and the disagreement among raters. And reliability is of course a prerequisite (but no guarantee) for the validity of the assessment of writing proficiency (see also Bachman, Lynch and Mason (1995) for a discussion of tension between the validity of direct assessment and reliability). The solution in most (national) assessments seems to be to apply multiple writing assignments, and to have students' performances rated by panels of expert readers. This solution is quite expensive, especially in the context of large-scale assessments. It is therefore interesting to see whether alternatives are available that can be used in assessment studies – alternatives with respect to the writing assignment and alternatives with respect to the reading procedures.

The different kinds of writing assignments, especially the ones that do not require readers (e.g., multiple-choice assignments), and the qualities of these assignments have been investigated in several studies (Veal and Tillman, 1971; Moss, Cole and Khampalikit, 1982; Quellmalz, Boodoo and Garlinghouse, 1983; Park, 1987; Capell and Chou, 1982; Schoonen, 1993). These different types of assignments will not be the main focus of this article, although the assignment-type can be considered an intervening variable in the study of the other aspect of the writing assessment, i.e., the rating of the writing performances.

This other aspect of the writing assessment, the 'quality' of the rating of the writing performances, has mainly been approached from the perspective of improving the scoring procedures: analytic and primary trait scoring instead of holistic scoring, so that the number of readers could be reduced. However, it is not only the scoring procedures that are important but also the expertise of the readers. This latter has been less frequently studied. It may be that the rating of certain writing tasks with certain scoring procedures can be performed by nonexpert (and less expensive) readers, or it may be that a (high) level of expertise is required no matter what kind of rating has to be done. Of course, one may expect the expertise of the reader to be relevant to the reliability of the rating. In order to come to reliable and valid ratings of essays, a reader should have a clear picture of what adequate writing looks like. The reader should at least have so-called 'domain knowledge' (cf., Alexander and Judy, 1988), in this case, knowledge of language usage, text organization and communicative adequacy. On the other hand, however, everyone who is a member of the language community and is a conceivable addressee of the written texts should be able to judge the quality of the texts. Does the text provide in an unambiguous way the information it
should provide? The conceivable addressee, i.e., the lay reader, is the more likely able to judge a written text when he or she can be expected to be able to perform the task quite easily him or herself. This could, for example, be the case with texts for young children.

Readers have to pretend to be a naive reader with the same knowledge as the intended readership of the text and, at the same time, they have to be alert to flaws in the text that could cause misinterpretations or miscommunication in a real-life situation. The question is whether lay readers are able to perform such a task and, if so, under which conditions they are able to do so. By conditions we mean kinds of assignment, kinds of scoring procedures and text qualities (or subscales) that have to be rated. For example, the task to rate simple, well defined types of text may require less expertise than does that of rating complex and ill-defined types of texts.

Some studies have been done on the role the reader's expertise plays in rating essays, the findings of which suggest that this expertise does affect the quality of the ratings. Diederich, French and Carlton (1961) studied the possible existence of 'schools of thought' among readers. Fifty-three readers rated 300 texts. The readers came from a variety of professional backgrounds and included social scientists, natural scientists, lawyers, writers/editors, business people and English (mother-tongue) teachers (7–10 readers per professional background). The median inter-reader correlation within the group of English teachers was the highest (0.41). Inter-reader correlation in the other professional groups ranged from 0.22 (business people) to 0.38 (lawyers), with a median of 0.28. Mother-tongue teachers seemed to agree more about the quality of texts than other readers.

McColly and Remstadt (1963, cited in McColly, 1970) experimented with four panels of four readers each. They observed that one panel deviated positively from the three other panels with respect to inter-reader agreement. Further study of the readers' characteristics revealed that the readers of the 'deviant' panel had an educational level which was 'substantially higher' than those of the other three panels.

Michael et al. (1980) compared the rating of experts, i.e., professors of English, with the rating of 'lay' readers, i.e., professors of other departments. Twenty expert and 20 lay readers read two sets (topic A and topic B) of 100 essays each. Their ratings were expressed on a four-point scale. Michael et al. conclude that the ratings of lay and expert panels (i.e., combinations of two readers) lead to similar mean scores for the sets. Expert readers agree somewhat more than do lay readers (inter-reader reliability topic A/B: 0.78/0.92 versus 0.75/0.78). Reading skills (i.e., expertise) affect reading reliability less than the topic. To evaluate the validity of the ratings,
Michael et al. compared them to three external criteria, two standard writing tests and the students' grade point average. The ratings of the experts turned out to be somewhat more valid than those of the lay readers, but again, the topic seems to affect the validity of the ratings to a much greater extent than does the assumed reading expertise.

Dunstan compared university readers to nonuniversity readers (Dunstan, 1959, cited in Gosling, 1966). The ratings of the university readers proved to be more stable (intrareader reliability: 0.94 versus 0.84) and showed greater agreement (inter-reader reliability: 0.60–0.89 versus 0.27–0.67).

More recently, Meuffels (1989) studied the rating reliability of eight language teachers, eight mathematics teachers and eight managers (the latter being used to conducting rating tasks on a regular basis). Meuffels showed that the scores (on a 11-point scale) of language teachers were more stable and showed greater agreement than did those of the maths teachers and the managers. The means of the correlations between a first and a second rating (i.e., intrareader reliability) were 0.76 (language teachers) and 0.64 (both maths teachers and managers). The means of the correlations between all possible pairs of readers within a panel (two out of eight) were 0.66 (language teachers), 0.46 (maths teachers) and 0.40 (managers). Meuffels replicated his own study. The replication supported his previous findings only to a certain extent. The ratings of the language teachers were no longer the most stable (0.71 versus 0.58 and 0.72), but they still showed the greatest agreement (0.50 versus 0.40 and 0.34). Meuffels also investigated the validity of the ratings. He found that the language teachers noticed many more linguistic errors in the essays than the maths teachers and the managers. The ratings of the language teachers were therefore considered more valid.

Finally, in the field of ESL some recent studies are of importance with respect to the issue of expertise in reading. Three recent studies in the field of rating (oral foreign/second) language performance in general (Bachman, Lynch and Mason, 1995; Brown, 1995; Lumley and McNamara, 1995) will be discussed in the final section of this article.

Cumming (1990) investigated whether six expert readers (i.e., teachers with at least seven years' experience in ESL composition instruction) and seven novice readers (i.e., student-teachers with no prior teaching experience) could discern ESL proficiency from writing proficiency. Both proficiencies, of course, are important aspects

---

1This 'loss' of validity might be due to the attenuation of the relationship between lay readers' judgements and the criterion, because of the lower reading reliability of the lay readers. Michael et al. do not refer to this possible explanation.
Cumming found that expert readers as well as novice readers were able to do so. However, the ratings for ‘content’ and ‘rhetorical organization’ of the novice readers and of the expert readers differed significantly from each other, but the ratings for ‘language use’ did not. The difference was established by comparing the mean scores for a set of 12 selected essays. Cumming also collected verbal reports from the raters. By classifying the decision-making behaviours of both groups of raters, he found some qualitative differences. Expert readers appeared to have ‘a much fuller mental representation of the “problem” of evaluating student compositions, using a large number of very diverse criteria, self-control strategies, and knowledge sources to read and judge students’ texts’ (Cumming, 1990: 43). However, within the groups of novice and expert readers there were sometimes large differences in the decision-making behaviours.

Weigle (1994) also collected verbal protocols of raters. The focus of her study was somewhat different, but still is of importance to us. Weigle investigated the effects of training on raters of ESL compositions. Training intends to clarify the goals and aims of the rating, and the interpretation of the criteria. Furthermore, it is expected to improve the expertise of the raters. Sixteen subjects, of whom eight had already some expertise and eight had not, had to score four selected compositions (representing two proficiency levels on two different prompts) for content, rhetorical control and language on a ten-point scale. Weigle found that the mean difference between pretest and post-test was 1.66 for the inexperienced raters and 0.77 for the experienced raters. This furthermore showed that extreme differences, i.e., three or more points difference, which could be traced to four inexperienced raters, were not equally distributed across the three subscales – rhetorical control being the most problematic. The protocols were collected from those four inexperienced raters. Weigle tried to trace the causes of these changes in rating by studying the protocols and she tried to relate these differences to the training the raters had received. However, it remains unclear whether raters who came to the same scores at pre- and post-test might not have changed their rating strategies and processes. By selecting these raters who showed one or more extreme differences, Weigle took the risk of focusing on cases or ratings that are not representative for regular ratings of inexperienced raters.

The previous studies show a relationship between reading reliability and reader’s expertise. Expert readers seem to be more reliable raters than lay readers: they are more stable (cf., Dunstan, 1959; Meuffels, 1989; Weigle, 1994) and show more agreement (cf., Dunstan, 1959; Diederich, French and Carlton, 1961; McColly and
Remstadt, 1963; Michael et al., 1980; Meuffels, 1989). Some studies showed that the validity of the ratings, as far as this could be established, of the experts is higher than that of the lay readers (cf., Michael et al., 1980; Meuffels, 1989).

The question is whether this difference in reading reliability (and validity)\(^2\) remains if certain aspects of the writing assignment are changed, such as the topic of the essays, the restrictiveness of the writing assignment and the rating instructions, the kind of text quality being rated, or the level of the students’ writing ability. Some studies already indicate that this may not be the case (Michael et al., 1980; Cumming, 1990; Weigle, 1994). Furthermore, the definition of expert and lay might influence the size of the ‘expertise effect’. The studies mentioned are quite specific in this respect; they all focused on very well educated ‘lay’ readers who – in most cases – were asked to give their general impressions of essays written by college students.

1 Context of the study

In the context of feasibility studies for a national assessment of educational progress in The Netherlands, it was concluded that the assessment of writing ability needed further study (Wesdorp et al., 1986). These studies should focus on the kinds of tasks to be used, either free-writing assignments or multiple-choice tests, or tasks somewhere in between these two extremes. Related to this choice is of course the question of reading reliability: what kind of instructions should readers be provided with? Instructions to give a general impression or instructions to give detailed scores? Limiting the writers’ freedom in writing and the readers’ freedom in rating usually makes things easier and often more reliable (and less valid!). With simple assessment procedures the scoring could be performed by computers or secretaries (nonprofessional lay readers), making the assessment of the writing performance far less expensive.

The above assumptions were the subject of a study into the assessment of writing ability (Schoonen, 1991; 1993) which experimented with different kinds of tasks and scoring. As part of that study, preliminary comparisons were made between lay readers and professional readers performing different kinds of rating tasks. In this context, lay readers had to be ‘real’ lay readers, persons who were of the educational level of secretaries.

This article focuses on the role the expertise of the readers plays

\(^2\)The focus of this study is on reading reliability. Validity of course should be the ultimate criterion to evaluate the quality of ratings. However, reliability is a necessary condition to reach validity. In this study validity is only touched upon.
in rating different kinds of text, elicited with different kinds of writing assignments. Three studies will be discussed, each of which seeks to investigate whether lay readers, i.e., readers of average education (secretaries or clerks) with no specific linguistic or educational schooling, rate essays with a degree of reliability comparable to that of experts or professional readers, i.e., (language) teachers. The three studies were all part of the major project mentioned above. We refer to them as different studies because the data (of the expert ratings) were collected in three different studies with different samples of students and different designs. These differences will be dealt with later on. First, we return to the parameters that may influence the comparison of lay and expert readers.

2 Aspect or subscale

Lay readers might be expected to be well capable of judging certain qualities of a text (e.g., Content or Mechanics) reliably, because they have strong intuitions or adequate knowledge about the aspect (subscale) in question. On the other hand, they may be incapable of doing so with other qualities (e.g., Usage or Style), since they have no knowledge or explicit criteria at their disposal. In all three studies we compare lay and expert readers with respect to their ratings of two different text qualities, one of which we consider relatively easy to judge (Content) and one we consider more difficult (Usage). These were the two text characteristics the national assessment was aiming at.

3 Task and topic

The difficulty of the rating task also depends on the extent to which the writing assignments are restrictive or free. Writing assignments which leave the topic to the writer's choice elicit very diverse essays, which complicates the rating task. More restrictive writing assignments for well specified communicative situations, or even revision tasks, elicit a limited range of 'responses' which simplifies the rating task.

In each of the three studies a different kind of writing assignment is used to elicit the texts from the students. Of each kind of writing assignment, three parallel versions were available only differing in the topic the students had to deal with (see also 'Methods', below).

4 Scoring

Furthermore, restrictive writing tasks allow the use of restrictive rating instructions. When a restricted range of 'written responses' can
be expected, detailed scoring guides can be developed. In free-writing assignments such scoring guides can hardly be imagined. Restrictive rating instructions may increase reading reliability even more. Therefore, different scoring procedures are used with the different writing assignments (see also ‘Methods’, below).

In sum, the more restrictive the writing and rating task and the easier the text quality to be rated, the smaller the difference we can expect in reading reliability between lay and expert readers. In the studies we report on here we collected ratings of lay and expert readers on two facets of the texts (Content and Usage), following two kinds of scoring instructions; there were four different kinds of writing assignments, each available for three topics. Each of the three datasets covers a part of the contrasts mentioned, so each dataset will be analysed separately (see also ‘Methods’, below).

5 Ways to compare

Comparisons of lay and expert readers – in the first place – involved a comparison of the perceived average quality of the texts, i.e., a comparison of the (mean) scores awarded by the expert and lay readers.

Further comparisons are made by comparing the reliability of the lay and expert readers. So far, we have referred to reliability in general terms. Reliability can be defined more specifically as the ratio of true score variance and total variance, i.e., true score variance and error variance (see Equation 1). In a factor analytic design the reliability (\(\rho\)) is estimated by

\[
\rho_{xx} = \frac{\lambda^2}{\lambda^2 + \theta}
\]

where \(\theta\) represents the error variance of the scores, and \(\lambda^2\) refers to the true score variance of the scores (cf., Jöreskog, 1971; Fleishman and Benson, 1987).

Equal reliability of the lay and expert readers only implies an equal ratio of true and error variance. This can be achieved in the context of completely different amounts of score variance. We have compared the amount of true score variance and the amount of error variance in the ratings of the lay and expert readers by evaluating the tenability of certain measurement models (Fleishman and Benson, 1987) (see also ‘Analysis’, below).

For the validation of the ratings of the lay and expert readers an external independent criterion would preferably be available. Such a criterion was not available for these studies. However, another indication of the validity of their ratings could be obtained. We examined
whether lay readers discriminate between Content and Usage in the way expert readers do. When the expert and lay readers rate the same features as Content or as Usage, the correlation between Content and Usage should be the same for both lay and expert readers.

II Methods

The writing samples and the ratings of lay and expert readers are compiled in three studies. These studies were pilot and feasibility studies for a national assessment in The Netherlands. Since we had to deal with data from different studies, the design for the comparisons between the lay and expert readers differs per study. We will explain the design for each comparison (see also ‘Materials’, below).

1 Subjects

Both writers and readers are involved in the studies. The writers are students approaching the end of primary education, grade 6 (mean age 12 years). The samples of students can be considered random samples of student writing. In this respect the proficiency level is what readers in a national assessment may expect.

The expert readers are experts because of their educational level and profession. They had linguistic and/or educational schooling and some experience in education as a teacher. Furthermore, the experts had previously participated in parts of the feasibility study for the national assessment. In sum, they were the kind of readers who will have to do the reading of writing assessments.

The lay readers were less well defined as a group. Essential parameters of their background should be that they have no specific linguistic or educational schooling, and that they have some administrative schooling and work as a secretary or similar job. The lay readers were recruited through a job centre and a school for commercial and administrative education. This is postsecondary education for the lower and average streams of secondary education.

The number of readers and writers will be specified per study.

2 Materials

a Tasks: There is a continuum of task types, ranging from a completely free choice of topic to multiple choice, from which four tasks can be distinguished that are used in our studies. These tasks were considered for use in the national assessment and therefore had to be studied. The most free task is called the ‘specified task’. It consists of a fully specified rhetorical situation upon which the students are
to base the writing assignment, but it leaves them free to generate and select information for their text. The second is the ‘structured task’. This task not only consists of rhetorical specifications but also of guidelines by which the students can gain a grasp on the goal of the task and generate and organize the required text. Finally, two semi-direct writing tasks provide students with texts that supposedly fulfil the requirements of the rhetorical situation, but need to be revised. These tasks are based on the interlinear tasks as first developed by Godshalk, Swineford and Coffman (1966). In one of these two tasks the parts of the text that need revision are underlined; in the other it is up to the student to decide which parts need to be revised. These semi-direct tasks are further referred to as the ‘interlinear revision task’ and the ‘interlinear revision task with marked errors’.

**b Topics:** Of the four kinds of writing assignment three versions were available, i.e., assignments with three different rhetorical specifications. For convenience, these different specifications will be referred to as topics. Topic A requires a persuasive essay on the question: ‘Should the killing of animals for fur be permitted?’ The stimulus for this essay was a fictitious newspaper article. The essay was meant to be ‘published’ in a school newspaper to inform the author’s peers of the writer’s opinion in these matters. Topic B asks for ‘a description of a holiday farm’, again for a school newspaper. The intended audience were the children of the school (but a different grade) who did not participate in the school holidays. A stimulus was provided by two drawings of the holiday farm, of which one was a plan. With topic C students are expected to write the instructions for a puzzle which consists of a maze. The intended readership were again the readers of the school newspaper. The stimulus was the maze puzzle and an explanation of how to proceed.

**c Reading tasks:** The reading procedure used with the specified writing task is rating on the basis of a scale with benchmark essays (cf., Study 1). A scale was constructed (in separate studies) for each of the three topics and for each of two text qualities (Content and Usage). A scale consists of five benchmark essays indicating average quality (score: 100) and one and two standard deviations below and above average (scores: 70, 85, 115 and 130). The reader has to compare the subject essay with the benchmark essays and has to decide which score represents the quality of the essay best, given the benchmarks. The scales allow all possible scores within and outside the
marked range. It is a way of impressionistic scoring, but it is supported by the scale of benchmark essays to avoid shifts in the standards the reader uses. Along with the scales a ‘definition’ of the aspect (either Content or Usage) was provided as well as a description of what a good essay would look like.

The structured task was scored according to scoring guides (cf., Study 2). For Content, readers were required to check for relevant and necessary propositions; these propositions were defined in the instructions and were of course different for the three topics. For Usage readers had to make counts of errors in grammar, vocabulary/idiom and style. These instructions were essentially the same for all three topics, but the examples provided were derived from the topic they had to rate.

Finally, expert and lay readers were asked to evaluate whether the students’ revisions ‘elicited’ by both the interlinear tasks improved the given text or not (cf., Study 3). Readers were given a list of weak spots in the submitted texts that ‘asked for’ revision. For each spot a number of possible improvements (and nonimprovements) were given in the scoring guide. For each topic, a student had to revise two or three texts per aspect. Different booklets were made for texts that needed to be revised with respect to Content and with respect to Usage, i.e., students knew which aspect to pay attention to.

Limitations in time and budget did not allow the investigation of all possible combinations of tasks and scoring procedures. But not all combinations make sense (e.g., general impression scoring of multiple-choice writing tests). However, since task and scoring are not completely crossclassified, conclusions about the influence of task and scoring procedure are contaminated.

3 Procedure

For every reading task the readers were trained by one of the researchers. The training consisted of an explanation of the specifications of the writing task, the criteria for the scoring and the way to proceed. After that a sample of essays was scored and the scores were discussed, and additional explanations were given. Practising continued until the readers felt confident about the interpretation of the criteria and the rating task. The readers rated the essays and revisions (in the interlinear tasks) independently, and the rating for Content and Usage was done separately to avoid contamination of the ratings.
**Analysis**

First, the average quality of the students’ essays and revisions according to the lay and expert readers are compared. A comparison of the strictness of both kinds of readers is performed by testing the differences in mean score of the lay and expert readers (t-test for dependent samples). Analyses are conducted for students with complete data (‘list-wise deletion’).

Secondly, the reading reliability, i.e., the amounts of true and error variance (see ‘Introduction’, above) of panels of lay and expert readers, is contrasted in covariance structure models (Jöreskog and Sörbom, 1989). These analyses allow us to test whether or not the lay and expert panels are parallel, tau-equivalent or congeneric (Jöreskog, 1971; Fleishman and Benson, 1987). Panel scores are considered parallel when they have an equal amount of true score variance ($\chi^2$) and an equal amount of error variance ($\theta$) as well, and thus an equal amount of total variance. Parallelism can also be more strictly defined by requiring the means to be equal too (cf., Lord and Novick, 1968). The term parallel is used here to refer to a situation of equal true score variance, and equal error variance (cf., Jöreskog, 1971; Fleishman and Benson, 1987). The panel scores are tau-equivalent when the scores show an equal amount of true score variance, but different error variance. The scores are considered congeneric when they stem from one latent variable, regardless of the true score variance and error variance.

Furthermore, we analysed whether the lay and expert readers discriminate between Content and Usage in the same way. We tested whether the correlation between Content and Usage is the same for the scores of the lay panel as it is for those of the expert panel. As stated before, this can be considered as an indication of equal validity of the lay and expert scores. Independent evaluation of the validity of the expert and the lay scores was not possible in these studies.

The tenability of a model of parallelism, tau-equivalence or congenericity of the expert and lay panel scores, and of equal correlation between Content and Usage, can be evaluated with LISREL (Jöreskog and Sörbom, 1989). Since it could not be assumed that the panel scores have a multivariate normal distribution, models are fitted according to the generalized least squares (GLS) method. Consequently, the models cannot be tested statistically. We used Fornell and Rust’s procedure (1989) to determine which of the proposed models is the most adequate. This procedure attributes a probability of accuracy to every model of a given set of concurrent models. The
probability depends on the fit of the model and the number of estimated parameters. Within a set of concurrent models these probabilities sum up to 1.

III Results

1 Study 1: rating of specified tasks with the use of scales with benchmark essays

In this study we compare the ratings the essays received from expert readers in a national assessment study with the lay ratings we collected ourselves. A random selection of about 125 essays per topic that were already rated by three experts for Content and by two for Usage were rated by four lay readers. In order to establish the same panel size for Content and Usage, selected lay readers were eliminated from the study. In the national assessment, the essays on each topic were written by different samples of students and therefore in the analysis each topic is treated as a different sample.

a The average quality: Table 1 summarizes the average quality of the essays as perceived by the expert and the lay panel. The table

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>95.7</td>
<td>94.8</td>
<td>11.4</td>
<td>14.8</td>
<td>109</td>
<td>109</td>
<td>2</td>
<td>93.4</td>
</tr>
<tr>
<td>B</td>
<td>91.3</td>
<td>86.9</td>
<td>12.0</td>
<td>14.9</td>
<td>119</td>
<td>119</td>
<td>2</td>
<td>93.8</td>
</tr>
<tr>
<td>C</td>
<td>94.3</td>
<td>98.7</td>
<td>15.0</td>
<td>14.7</td>
<td>119</td>
<td>119</td>
<td>2</td>
<td>105.1</td>
</tr>
</tbody>
</table>

Note: The difference in mean panel scores between lay and expert readers is statistically significant (t-test for related samples, α = 0.05 corrected for capitalization on chance according to Holm, 1979).

3 Initially five readers were recruited to do the job, but one withdrew from the study at an early stage because he felt too uncertain of his rating skills.

4 The final results of Table 3 were checked by estimating the panel reliability of all possible combinations of three and two lay readers, respectively. These 24 (partly dependent) replications showed similar results to the initial analyses.
shows a number of differences in the stringency with which the panels rated the essays. For Content scores there are two (statistically significant) differences: the lay readers are more lenient with topic B, and the expert readers with topic C. With respect to Usage, the lay panel seems to be more lenient than the expert panel (topics B and C). The lay and expert panel agree on the average quality of topic A essays (Content and Usage).

b Reader reliability: In order to test the aforementioned types of 'similarity' in the panels, i.e., parallelism, tau-equivalence and congenerity for either Content or Usage ratings, or both, we fitted several measurement models. As every topic in this study corresponds to a different sample of students, the models are fitted in three samples (topics A, B and C) simultaneously (multisample analysis). The analysis is therefore done with three matrices of order 4: two lay panel scores (Content and Usage) and two expert panel scores. Figure 1 illustrates the general model for one sample. Table 2 lists the models and the restrictions imposed on the parameters. This table shows a large difference in fit ($\chi^2/df$ and $gfi$) between the models (I to VI and models VII and VIII). Assuming that one of the eight models is correct, it is most probably model VII (cf., pfr, Fornell and Rust, 1989). This means that the expert and lay panel are neither parallel nor tau-equivalent in their ratings of the essays. The differences in standard deviation (cf., Table 1) made parallelism a very unlikely possibility. The expert readers tend to use a larger range of scores. Expert and

![Figure 1](https://example.com/figure1.png)  
*Figure 1*. General model for the comparison of the lay and expert panels: rating of the specified tasks with the use of scales with benchmark essays
Table 2 Models and their fit ($\chi^2/df, \text{gfi}$) and the degree of likelihood according to Fornell and Rust ($p_r$): specified tasks rated with scales of benchmark essays ($N_a = 109$, $N_b = 119$, $N_c = 119$) (cf., Figure 1)

<table>
<thead>
<tr>
<th>Model</th>
<th>Restrictions</th>
<th>Characteristics</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2/df$</th>
<th>gfi*</th>
<th>$p_r$</th>
</tr>
</thead>
</table>
| I     | $\lambda_1 = \lambda_2$
|      | $\lambda_3 = \lambda_4$
|      | $\theta_1 = \theta_2$
|      | $\theta_3 = \theta_4$
|      | panels are parallel for both Content and Usage | 96.3  | 17  | 5.7 | 0.859 | 0.000 |
| II    | $\lambda_1 = \lambda_2$
|      | $\lambda_3 = \lambda_4$
|      | $\theta_1 = \theta_2$
|      | $\theta_3 = \theta_4$
|      | panels are parallel for Content, tau-equivalent for Usage | 85.3  | 14  | 6.1 | 0.875 | 0.000 |
| III   | $\lambda_1 = \lambda_2$
|      | $\lambda_3 = \lambda_4$
|      | $\theta_1 = \theta_2$
|      | $\theta_3 = \theta_4$
|      | panels are parallel for Usage, tau-equivalent for Content | 88.2  | 14  | 6.3 | 0.872 | 0.000 |
| IV    | $\lambda_1 = \lambda_2$
|      | $\lambda_3 = \lambda_4$
|      | $\theta_1 = \theta_2$
|      | panels are tau-equivalent for both Content and Usage | 77.1  | 11  | 7.0 | 0.888 | 0.000 |
| V     | $\lambda_1 = \lambda_2$
|      | panels are tau-equivalent for Content, congeneric for Usage | 43.5  | 8   | 5.4 | 0.937 | 0.000 |
| VI    | $\lambda_3 = \lambda_4$
|      | panels are tau-equivalent for Usage, congeneric for Content | 54.1  | 8   | 6.8 | 0.921 | 0.000 |
| VII   | —
|      | panels are congeneric for both Content and Usage | 8.5   | 5   | 1.7 | 0.988 | 0.504 |
| VIII  | $\text{cov}(e_3,e_4) \neq 0$
|      | similar to model VII allowing extra correlation between Content and Usage for the lay readers | 6.6   | 4   | 1.6 | 0.991 | 0.496 |

Note:
* gfi is the mean of three goodness-of-fit indices. LISREL computes a goodness-of-fit index per sample.

lay readers' scores not only differ in their total (observed) variance but also in the 'true score' variance (i.e., the rejection of model IV of tau-equivalence).

Allowance for extra correlation between Content and Usage (as in model VIII) does not improve the description of the data compared to a model without this extra correlation (model VII), so we can conclude that the extent to which the lay and expert panels discriminate between Content and Usage is more or less the same. Estimated true correlation is 0.81 (correlation between Content and Usage was assumed to be the same across the three topics).

Finally, we want to draw attention to the differences between Content and Usage. Restrictions imposed on true and error variance in the Content scores affect the fit of the model to the data to a much lesser extent than do the same restrictions when imposed on true and
error variance in the Usage scores. This may be interpreted as the lay panel and the expert panel seeming to be more ‘similar’ in their Content ratings than in their Usage ratings.

Table 3 gives the estimated panel reliabilities (see Equation 1), estimated within model VII. The reliability of the lay panel for Usage is systematically lower than that of the expert panel. The reliability for Content differs (substantially) for topic B only. This topic is rated more reliably by the expert panel than by the lay panel.

<table>
<thead>
<tr>
<th></th>
<th>Content ((k=3))</th>
<th>Usage ((k=2))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert</td>
<td>Lay</td>
</tr>
<tr>
<td>Topic A ((n=109))</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Topic B ((n=119))</td>
<td>1.00*</td>
<td>0.70</td>
</tr>
<tr>
<td>Topic C ((n=119))</td>
<td>0.88</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Note: * This reliability was estimated at 1.046 due to a slightly negative estimate of the corresponding error variance.

2 Study 2: rating of the structured tasks with the use of scoring guides

Along the same line of reasoning, we compared the ratings using the scoring guides of an expert panel with those of a lay panel and with the texts being elicited by structured writing tasks. The essays for the three topics were written by one sample of the same students, different students from Study 1.

Although we set out to obtain ratings from five lay and five expert readers (all to be collected by ourselves), only four lay readers actually started the job (cf., note 3) and one of those four did not finish this job (because of personal reasons). In order to establish same panel size, the ratings of two randomly selected experts were eliminated from the study.

a The average quality: Table 4 summarizes the average essay ratings for both the expert and lay panels. Four of the six comparisons of mean panel scores of the expert and lay readers show a statistically significant difference. In these cases the lay panel is stricter in its
Table 4  Structured task with scoring guides: mean panel score ($M$) and standard deviation ($s$) for lay and expert readers. $N$ is the number of texts, and $k$ is the number of readers.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Content</th>
<th>Usage**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>$M$</td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>3.7</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>4.5*</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.2</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>4.3</td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>3.5*</td>
</tr>
</tbody>
</table>

Notes:
* The difference in mean panel scores between lay and expert readers is statistically significant ($t$-test for related samples, $a = 0.05$ corrected for capitalization on chance according to Holm, 1979).
** Scoring of usage is an error count. Therefore, low numbers indicate good writing.

judgement than the expert panel. They score fewer Content elements and more Usage errors.

b Reader reliability: In order to establish the reading reliability of lay and expert panels we once more tested the eight models for 'similarity'. As all three topics were assigned to one and the same sample of students, a one-sample analysis was conducted with a matrix of order 12: 3 (topics) $\times$ 2 (levels of expertise) $\times$ 2 (aspects, Content and Usage). This analysis is suitable for evaluating the same restrictions as listed in Table 2. Figure 2 depicts the model. Table 5 describes the eight models and their fit to the data. The most restrictive model (I) already shows a reasonable fit ($\chi^2/df$ and $gfi$). However, releasing the restriction of parallelism improves the fit of the model to some extent. Again, the restriction seems more tenable for Content than for Usage (see also the differences in standard deviation of the lay and expert panel scores for Content and Usage, Table 4). According to the Fornell and Rust procedure, model IV, which assumes tau-equivalent panels, is the most acceptable model. On the whole differences between the models seem to be quite small.

As in Study 1, the correlation between Content and Usage is considered to be the same for each topic ($\psi_1 = \psi_2 = \psi_3 = \psi$, see Figure 2). The correlation between Content and Usage for the expert and lay panels may be considered the same, since model VIII fails to produce
better results than model VII. The estimated correlation between Content and Usage is 0.10.

We used the most appropriate model (IV) to estimate the panel reliabilities which are shown in Table 6. Small differences are to be expected for Content, given the small difference in fit between model IV and II. However, in rating topic A the lay panel seems to be more reliable than the expert panel. Topics B and C show smaller differences, alternating in favour of the expert and the lay panels.

Considering topics A and B, Usage is clearly rated more reliably by the expert than by the lay panel. Topic C shows no difference. Topic C essays are somewhat shorter than topic A and B essays. It might be that with topic C experts had little opportunity to show their skill in rating essays. The relatively few usage errors the children made (see Table 4) support this view.
Table 5  Models and their fit ($\chi^2$/df, gfi) and the likelihood according to Fornell and Rust ($p_n$): structured task with scoring guides ($N=99$) (cf., Figure 2)

<table>
<thead>
<tr>
<th>Model</th>
<th>Restrictions</th>
<th>Characteristics</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2$/df</th>
<th>gfi</th>
<th>$p_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$\lambda_1 = \lambda_2$ $\theta_1 = \theta_2$ $\lambda_3 = \lambda_4$ $\theta_3 = \theta_4$ $\lambda_{12} = \lambda_{12}$ $\theta_{11} = \theta_{12}$</td>
<td>panels are parallel for both Content and Usage</td>
<td>74.5</td>
<td>53</td>
<td>1.4</td>
<td>0.873</td>
<td>0.002</td>
</tr>
<tr>
<td>II</td>
<td>$\lambda_1 = \lambda_2$ $\theta_1 = \theta_2$ $\lambda_3 = \lambda_4$ $\theta_3 = \theta_4$ $\lambda_{12} = \lambda_{12}$ $\theta_{11} = \theta_{12}$</td>
<td>panels are parallel for Content, tau-equivalent for Usage</td>
<td>58.3</td>
<td>50</td>
<td>1.2</td>
<td>0.901</td>
<td>0.317</td>
</tr>
<tr>
<td>III</td>
<td>$\lambda_1 = \lambda_2$ $\theta_1 = \theta_2$ $\lambda_3 = \lambda_4$ $\theta_3 = \theta_4$ $\lambda_{12} = \lambda_{12}$ $\theta_{11} = \theta_{12}$</td>
<td>panels are parallel for Usage, tau-equivalent for Content</td>
<td>68.0</td>
<td>50</td>
<td>1.4</td>
<td>0.884</td>
<td>0.002</td>
</tr>
<tr>
<td>IV</td>
<td>$\lambda_1 = \lambda_2$ $\lambda_3 = \lambda_4$ $\lambda_{12} = \lambda_{12}$</td>
<td>panels are tau-equivalent for both Content and Usage</td>
<td>52.3</td>
<td>47</td>
<td>1.1</td>
<td>0.911</td>
<td>0.319</td>
</tr>
<tr>
<td>V</td>
<td>$\lambda_1 = \lambda_2$ $\lambda_3 = \lambda_4$ $\lambda_5$</td>
<td>panels are tau-equivalent for Content, congeneric for Usage</td>
<td>46.5</td>
<td>44</td>
<td>1.1</td>
<td>0.921</td>
<td>0.287</td>
</tr>
<tr>
<td>VI*</td>
<td>$\lambda_1 = \lambda_2$ $\lambda_3 = \lambda_4$ $\lambda_5$</td>
<td>panels are tau-equivalent for Usage, congeneric for Content</td>
<td>50.9</td>
<td>46</td>
<td>1.1</td>
<td>0.914</td>
<td>0.032</td>
</tr>
<tr>
<td>VII*</td>
<td>$\lambda_1 = \lambda_2$ $\lambda_3 = \lambda_4$ $\lambda_5$</td>
<td>panels are congeneric for both Content and Usage</td>
<td>45.3</td>
<td>43</td>
<td>1.1</td>
<td>0.925</td>
<td>0.026</td>
</tr>
<tr>
<td>VIII*</td>
<td>$\text{cov}(e_2,e_6) + \text{cov}(e_4,e_{10}) = \text{cov}(e_6,e_{12}) \neq 0$</td>
<td>as model VII allowing extra correlation between Content and Usage for the lay readers</td>
<td>44.4</td>
<td>42</td>
<td>1.1</td>
<td>0.925</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Note: * In models VI–VIII two error variances were estimated as being negative and the algorithm did not converge within 250 iterations. Both were then fixed at zero. This leads to two more degrees of freedom. In computing the likelihood ($p_n$) we used the number of free parameters of the intended model.

3 Study 3: rating of interlinear revision tasks with the use of scoring guides

Study 3 deals with the rating of revisions students made in texts about topics A, B and C. The texts were ‘constructed’ to represent either a number of Content flaws or a number of Usage errors. With the unmarked version of the interlinear revision task the students were required to search for and to correct unmarked errors; in the marked version they only had to revise the (marked) errors. One sample of students participated in revising texts with the unmarked errors on all
three topics. Another sample of students revised texts (three topics) with marked errors.

When the data were collected it was found that, despite an earlier pilot study, Content revision clearly required more time than the students were allowed, both for marked error and unmarked error texts. Numerous students were unable to finish their revision task for Content in time, as can be seen from the small sample sizes in Table 7. Therefore, only the mean panel scores will be compared for both aspects, Content and Usage; comparisons of reader reliability will be restricted to Usage solely.

As it was expected that the revisions were rated far more reliably than the texts, only three expert and three lay readers were invited to rate the interlinear revision tasks.

**a The average quality:** Table 7 illustrates some differences in the perceived quality of the revisions the students made. Of the 12 comparisons, six show a difference between the expert and the lay panels, five of which relate to Usage. All statistically significant differences in panel means indicate that, again, lay readers tend to be more stringent in their judgements than expert readers (cf., Study 2).

**b Reader reliability:** Reading reliability was established for Usage solely. The two types of interlinear revision tasks (with and without marked errors) were assigned to two different samples of students, so we have to conduct a two-sample analysis. Figure 3 depicts the general model for one sample and shows that all students within a sample made revisions for all three topics. The models to be tested are fitted simultaneously for the two samples (i.e., students revising the marked and those revising the unmarked errors). Therefore, two covariance matrices of order 6 (three topics by two panels) were analysed. As it was necessary to drop the Content scores, there are of course fewer models to fit. For the sake of convenience, the same
Table 7 Interlinear revision tasks: mean panel score ($M$) and standard deviation ($s$) for lay and expert readers. $N$ is sample size, and $k$ is the number of readers.

<table>
<thead>
<tr>
<th>Interlinear revision task</th>
<th>Content</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>$M$</td>
</tr>
<tr>
<td><strong>Topic A</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>1.9</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Topic B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>1.3</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Topic C</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>1.4</td>
</tr>
</tbody>
</table>

**Interlinear revision task with marked errors**

<table>
<thead>
<tr>
<th>Interlinear revision task with marked errors</th>
<th>Content</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>$M$</td>
</tr>
<tr>
<td><strong>Topic A</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>4.0</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>4.0</td>
</tr>
<tr>
<td><strong>Topic B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>1.9</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>2.0</td>
</tr>
<tr>
<td><strong>Topic C</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay</td>
<td>3</td>
<td>2.2</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>2.1</td>
</tr>
</tbody>
</table>

**Note:**

* The difference in mean panel scores between lay and expert readers is statistically significant ($t$-test for related samples, $\alpha_p = 0.05$ corrected for capitalization on chance according to Holm, 1979).

Model labels are used in Table 8 as in Tables 2 and 5. Table 8 summarizes the fit of the models to the data. Although the most restrictive model (I) has a reasonable fit, this model of parallelism (model I) and the model of tau-equivalence (model IV) show a poorer fit than the model of congenerity (model VII), even if we take the difference of degrees of freedom into consideration ($p_{fr}$).

In concluding these analyses it was not possible to compare the lay and expert panels in their discrimination between Content and Usage. Instead, we compared the correlations between Usage scores of the three topics.

A model that allows the correlations between the Usage scores of the three topics to differ for the expert and lay panel does little if anything to improve the fit of the model (model VIII versus model VII). Model VII seems to be the most likely model ($p_{fr}$). This model has been used to estimate the reliabilities of both panels. They are...
The general model for the comparison of the lay and expert panels: rating of the interlinear revision tasks with the use of scoring guides (Usage only)

![Diagram](chart.png)

**Table 8** Models and their fit ($\chi^2/df$, gfi) and the likelihood according to Fornell and Rust ($p_{fr}$): interlinear revision tasks for Usage, with and without marking of the errors ($N_{without} = 85$, $N_{with} = 96$) (cf., Figure 3)

<table>
<thead>
<tr>
<th>Model</th>
<th>Restrictions</th>
<th>Characteristics</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2/df$</th>
<th>gfi*</th>
<th>$p_{fr}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$\lambda_1 = \lambda_2$ $\theta_1 = \theta_2$ $\lambda_3 = \lambda_4$ $\theta_3 = \theta_4$ $\lambda_5 = \lambda_6$ $\theta_5 = \theta_6$</td>
<td>panels are parallel</td>
<td>66.7</td>
<td>27</td>
<td>2.5</td>
<td>0.875</td>
<td>0.009</td>
</tr>
<tr>
<td>IV</td>
<td>$\lambda_1 = \lambda_2$ $\lambda_3 = \lambda_4$ $\lambda_5 = \lambda_6$</td>
<td>panels are tau-equivalent</td>
<td>47.3</td>
<td>21</td>
<td>2.3</td>
<td>0.913</td>
<td>0.390</td>
</tr>
<tr>
<td>VII</td>
<td></td>
<td>panels are congeneric</td>
<td>34.8</td>
<td>15</td>
<td>2.3</td>
<td>0.937</td>
<td>0.495</td>
</tr>
<tr>
<td>VIII</td>
<td>$\text{cov}(e_2,e_4) = 0$ $\text{cov}(e_2,e_6) = 0$</td>
<td>as model VII allowing extra correlation between the topics for the lay readers</td>
<td>33.9</td>
<td>13</td>
<td>2.6</td>
<td>0.938</td>
<td>0.106</td>
</tr>
</tbody>
</table>

* gfi is the mean of two goodness-of-fit indices. LISREL computes a goodness-of-fit index per sample.

As is shown in the table, panel reliabilities are relatively high – all above 0.90. The differences in reliability are very small for the interlinear revision tasks without marked errors and only slightly larger for the revision task with marked errors. It is remarkable that in rating this latter task the lay panel proved to be more reliable than the expert panel.
Table 9 Estimated panel reliability for the interlinear revision tasks with and without marking of the errors (solely Usage). Estimations in model VII, congeneric panels. N is sample size, k the number of readers.

<table>
<thead>
<tr>
<th></th>
<th>Interlinear revision task with unmarked errors (N=85, k=3)</th>
<th>Interlinear revision task with marked errors (N=96, k=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert</td>
<td>Lay</td>
</tr>
<tr>
<td>Topic A</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Topic B</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Topic C</td>
<td>0.96</td>
<td>1.00*</td>
</tr>
</tbody>
</table>

Note: *This reliability was estimated at 1.001 due to a slightly negative estimation of the corresponding error variance.

IV Conclusion and discussion

In general, we can conclude that lay readers and expert readers are not interchangeable. Of 24 comparisons between the lay and expert panels with respect to their average scores 14 demonstrate a difference between the lay panel score and the expert panel score. The differences were found in all three parts of this study. Nine of these 14 differences were found in the assessment of Usage. In 11 cases the lay readers were stricter than the expert readers; the three exceptions were found in the first study about the use of scales of benchmark essays.

Differences in reading reliability were clearest in the use of essay scales (Study 1), especially with respect to the rating of Usage. The expert readers were more reliable than the lay readers. The use of scoring guides (Study 2) demonstrates a similar tendency, though less clearly. However, the rating of the revisions (Study 3) shows quite a different picture. Differences in reader reliability with respect to Usage are very small and seem to be in favour of the lay readers.

For Content the differences in reading reliability are smaller and alternately in favour of the lay and expert panels.

Correlations between the two subscales, i.e., Content and Usage, or topics proved to be independent of the expertise of the panel (models VIII were no better than models VII). Therefore, we can conclude that the extent to which the lay and expert panels discriminate between Content and Usage (and the three topics, Study 3) is more or less the same. The panel scores can, in this respect, be considered equally valid.
Practical implications

The findings of the studies mentioned in the introduction, which suggest that expert readers are more reliable than lay readers, need to be differentiated. Our analyses showed that the findings depend upon (the combination of) the kind of writing and rating task that are used and on the aspect of the text that has to be rated (in our case Content and Usage). These conclusions are in line with the recent findings of Brown (1995), who found that raters with different professional backgrounds did not differ in their overall rating of oral language proficiency, but did differ in the way which they perceived the assessment criteria and the way in which they applied the criteria.

We have, of course, not studied all possible crossclassifications of aspects, writing tasks and rating procedure, but just four combinations that were considered for national assessment purposes. And the samples of lay and expert readers were quite small, although we think they were representative or indicative for their 'populations'. Nevertheless, we consider that the study shows some tendencies that may be relevant in the selection of readers in (large-scale) writing assessment, and may guide future research into the role of expertise in reading reliability.

If one has to decide on the use of lay readers in (large-scale) writing assessment one can use different criteria for the lay readers to meet. One can use an absolute criterion which the lay readers have to meet, e.g., ‘they have to reach a panel reliability of 0.80’. One can also apply a relative criterion, namely, that the use of lay readers is a profitable possibility in so far as their reliability corresponds to their cheaper wages. For example, if the lay readers are three times less expensive than the experts, their panel size can be three times as large for the ‘same price’. The Spearman–Brown prophecy formula for homogeneous test lengthening can be used to calculate whether it is worth while employing lay readers. If one does not want to enlarge the panels, one can simply state that the lay readers only will be used when they are as reliable as expert readers. For all these three options the reliability estimates provide sufficient information (Tables 3, 6 and 9).

One can also use a somewhat different criterion, namely, the lay and the expert readers have to discern the same amount of true score variance. If this is the criterion one wants to apply, one can be satisfied with tau-equivalence or parallelism in the model fitting. This criterion is met by the rating of structured tasks by means of scoring guides. This criterion is not met by the rating of the revision tasks due to the lower true variance of the scores of the expert readers.

In sum, the criterion applied for the use of lay readers determines
the kind of tasks they can do. In general, the lay readers lag behind in the rating of Usage in writing tasks of grade-6 students that are relatively free (specified and structured task). The lay readers’ reliability in rating Content, and Usage in very restrictive tasks (revision of texts with and without marked errors), is comparable to the reliability of the experts.

Extra training of the lay readers seems to be of little use, since rater differences are to a large extent not susceptible to training (Lumley and McNamara, 1995).

2 Research context

Former studies (Dunstan, 1959, cited in Gosling, 1966; Michael et al., 1980; Meuffels, 1989; see also the introduction) also found differences between lay and expert readers. And in general, we can conclude that lay readers are less reliable in rating than expert readers. In our analyses, however, we showed that some factors have to be taken into consideration to differentiate this general conclusion from those studies. The types of writing and rating tasks and the aspects of the text to be rated are factors which should be taken into consideration when comparing lay readers’ and expert readers’ reliability. These factors might influence the difficulty of the rating task, as was stated in the introduction. For example, the required content elements in simple (functional) texts that our young students had to write are limited and more or less countable. This simplifies the rating tasks. Both lay and expert readers (as possible addressees themselves) can easily decide whether the text conveys the intended message. Usage, on the other hand, stands for the uncountable possible ways of expressing one’s (simple) message in language. In rating Usage, readers have to have explicit intuitions on appropriate language. Experts may be expected to have this kind of metalinguistic awareness, which guides their rating process. The more difficult the rating task is, the clearer the differences between expert and lay readers become.

In the same vein, the kind of response that is requested from the students is of course important. Are students free to write about the topic they want or do they have to respond to a strictly formulated assignment? Can they invent or formulate their text freely or is their response limited one way or the other? In the latter case the kinds of writing performances can be anticipated and the readers’ instructions can be confined to the anticipated performances; the rating task thus becomes easier, more reliable. In Study 3, for example, we used a relatively simple and well defined writing assignment, interlinear
revision tasks, which made it possible to give detailed scoring instructions. The detailed instructions may be very helpful to the lay readers and hinder the experts in showing their expertise.

Unlike the above-mentioned studies, we used young children's texts and revisions. Most studies concentrate on college or university students, whereas our study focused on primary-school children. This difference in the levels of writing ability of the writers might explain why in those studies experts were always the more reliable readers. In rating high-level texts, experts have more opportunity to show their (rating) skills; lay readers may have little knowledge of and experience with the standards for those texts and become insecure and unreliable raters. To test this hypothesis further, lay readers should be confronted with texts written by students of different levels of writing ability and the readers should undertake the writing assignment themselves.

Recent studies into rater characteristics in the field of (oral) performance testing (Bachman, Lynch and Mason, 1995; Brown, 1995; Lumley and McNamara, 1995) show interesting approaches for the study of the effects of rater characteristics or background which could be applied to future studies into the rating of compositions. Not only should reader background or expertise be a facet in such studies but also the type of writing assignment, the type of scoring procedure, the criteria or subscales that have to be applied and, last but not least, the level of writing performance of the students.

V References


Quellmalz, E.S., Capell, F.J. and Chou, C.P. 1982: Effects of discourse


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