Article

LANGUAGE TESTING

Language Testing
2015, Vol. 32(1) 83–100

DOI: 10.1177/0265532214542994
ltj.sagepub.com
SAGE

Effect of genre on the generalizability of writing scores

Renske Bouwer
Utrecht University, the Netherlands

Anton Béguin
Institute for Educational Measurement (Cito), Arnhem, the Netherlands

Ted Sanders
Utrecht University, the Netherlands

Huub van den Bergh
Utrecht University, the Netherlands

Abstract

In the present study, aspects of the measurement of writing are disentangled in order to investigate the validity of inferences made on the basis of writing performance and to describe implications for the assessment of writing. To include genre as a facet in the measurement, we obtained writing scores of 12 texts in four different genres for each participating student. Results indicate that across raters, tasks and genres, only 10% of the variance in writing scores is related to individual writing skill. In order to draw conclusions about writing proficiency, students should therefore write at least three different texts in each of four genres rated by at least two raters. Moreover, when writing scores are obtained through highly similar tasks, generalization across genres is not warranted. Inferences based on text quality scores should, in this case, be limited to genre-specific writing. These findings replicate the large task variance in writing assessment as consistently found in earlier research and emphasize the effect of genre on the generalizability of writing scores. This research has important implications for writing research and writing education, in which writing proficiency is quite often assessed by only one task rated by one rater.

Keywords

Generalizability theory, genre effect, rater effect, task effect, writing assessment

Corresponding author:
Renske Bouwer, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, Utrecht, 3512 JK, the Netherlands.
Email: i.r.bouwer@uu.nl

Assessment of writing proficiency is essential to writing education as well as writing research. For example, teachers want to know whether their students are able to write well-structured and understandable texts and researchers want to know whether their writing intervention was effective. As is the case for all performance assessments, the most appropriate way to assess writing proficiency is to have people write one or more texts (Huot, 1990b). It is, however, hard to generalize text quality scores to writing skills in general, as ratings of text quality do not only vary due to individuals' writing skills, but also due to characteristics of the measurement situation such as raters and tasks.

The effect of raters on text quality scores has been studied extensively in earlier research, but the effects of task are less well understood. Analyses of task effects have almost always been based on multiple tasks in one genre, or on single tasks within multiple genres, such as narrative and argumentative writing. Hence, topic and genre effects are confounded. It is still unclear whether generalization of writing scores is warranted across genres when writing assessments only include highly similar tasks (e.g., Van den Bergh, Maeyer, Van Weijen, & Tillema, 2012). On the other hand, effects of genre cannot be inferred when tasks differ both in genre and topic (Coffman, 1966; Veal & Tillman, 1971). In the current study, the effects of the writing task are further disentangled, allowing for a more valid interpretation of the results of writing assessments. The dual aim of this study is to (a) investigate and demonstrate the validity of inferences made on the basis of writing performance both within and across genres and (b) describe its implications for the assessment of writing proficiency.

One of the facets in the measurement that causes variance in text quality ratings, other than individuals' writing skills, is the rater. Raters are not always consistent in their judgments and they often disagree (Godshalk, Coffman, & Swineford, 1966; Schoonen, Vergeer, & Eiting, 1997). Rater variability may impact both absolute decisions (i.e., decisions concerning performance levels) and relative decisions (i.e., decisions concerning the ranking of students) that are made on the basis of writing performance. For instance, some raters are more strict than others (Weigle, 2002). Ratings of strict raters are consistently too harsh, in comparison to other raters or established benchmarks. When rater severity is not taken into account, student's writing performance will, in this case, be underestimated. Score variance may also be affected by interactions between rater and student. For instance, raters who differ in their interpretation and use of criteria for evaluating text quality (Eckes, 2008) or who differ in their expectation of, or involvement with, the writer (Wiseman, 2012) will rank order students' texts quite differently.

Another potential source of error in the assessment of writing is the writing task. Huang's meta-analysis (2009) showed that the two main sources of variation in performance scores are related to task characteristics. First, overall results indicated that roughly 10% of the variance is due to main task effects. This shows that average scores for performance quality differ between tasks, implying that tasks vary in level of difficulty. Second, approximately a quarter of the variance between performance scores is due to interaction effects between persons and task, implying that individuals do not perform consistently across tasks.

Hence, in the context of writing performance, characteristics of raters and tasks appear to play a significant role in the assessment. Ratings of text quality are always subjective to some extent. Moreover, text quality partly depends on the topic written about and the

genre written in, specified by the purpose for writing and the intended audience (Huot, 1990b). Decisions about students' writing performance are thus greatly influenced by characteristics of raters and tasks, implying that it is almost impossible to generalize to writing proficiency based on the quality of one written text, scored by one rater. Valid and reliable writing assessments should therefore include multiple tasks and raters. In addition, Lee and Kantor (2007) showed that the task facet explains more of the variability in the observed writing scores than the rater facet seems to do. Therefore, they argue, it is more efficient to increase the number of tasks in the assessment, than to increase the number of raters per task.

With this in mind, the question is how many tasks and raters are necessary for a reliable assessment of writing proficiency. Generalizability theory provides a framework for deciding upon the number of tasks and raters, given the multiple sources of measurement error (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991; for a basic introduction to this theory, see Bachman, Lynch, & Mason, 1995; Schoonen, 2005). Generalizability theory comprises a two-staged, multifaceted analysis. In the first stage, the so-called generalizability study (G-study), multiple sources of measurement error are disentangled and their variances are estimated. Based on estimates of variance components, the score generalizability can be described, reflecting the accuracy of generalizations made from the observed scores (i.e., text quality scores) to the "universe" scores (i.e., individuals' writing proficiency). In terms of generalizability theory, a universe score is the expected value of a person's observed scores over all observations to which a decision maker wants to generalize. In writing research, this universe of generalization includes all admissible raters and tasks, because one wants to generalize to writing performance across characteristics of raters and tasks (Gebril, 2009). In the second stage, the decision study (D-study), estimates of variance components obtained from the G-study are used to examine how variations in the assessment design affect score generalizability. The goal of the D-study is to choose the optimal number of tasks and raters that minimizes measurement error and consequently increases the generalizability of text quality scores to writing proficiency.

The optimal number of tasks and raters depends upon the purpose of the assessment or the decision one wants to make (Cronbach et al., 1972). In educational practices, writing scores are generally used to decide whether a student is able to write at a sufficient performance level. For absolute decisions about performance, measurement error includes both main effects of task and raters and task and/or rater effects that influence the ranking of students, that is, all interaction effects with students including random error. Writing scores are, however, at best interval scaled, rendering arbitrary means and variances of ratings (cf. Suppes & Zinnes, 1963). If means of individual tasks are arbitrary, so are differences in means between tasks. Hence systematic variability due to tasks has no relevance, implying that writing scores reflect relative group performance rather than performance pertaining to absolute standards. For relative decisions reflecting relative group performance, only task and/or rater effects that interact with persons contribute to measurement error.

Previous research applying generalizability theory to the assessment of writing showed that at least five tasks and three raters are necessary to make valid and reliable decisions about writing skills (Coffman, 1966; Schoonen, 2005, 2012, Van den Bergh

et al., 2012). Although the results of these studies seem to converge, they differ in the kind of tasks assigned to students. For instance, Van den Bergh et al. (2012) used highly similar argumentative tasks, whereas Schoonen (2005) included argumentative tasks as well as functional tasks (e.g., to give instructions, or to describe a route on a map). In some studies, tasks differed even more. Studies of Coffman (1966) and Veal and Tillman (1971), for instance, included tasks aiming at four different writing purposes, namely, describing, narrating, exposing and arguing. These tasks did also differ in the intended audience. For instance, in the study of Coffman (1966) students had to write personal essays as well as texts that were intended for another student.

Genre differences, shaped by the rhetorical situation in the writing task, are likely to lead to differences in the text that has to be produced. However, we do not know what the relationship is between genre and writing, because previous studies only included one task in multiple genres, or multiple tasks in one genre, thereby confounding the effects of genre and topic. Huot (1990b) indeed concluded in a literature review that task characteristics might elicit different writing quality, but that research on genre effects is inconclusive. In addition, Hoetker (1982) argued for a better operationalization of genre to study task variability in writing assessment.

Genre theorists and analysts, such as Swales (1990) and Bhatia (1993), have shown that texts that share the same communicative purpose and audience (i.e., texts in the same genres) are more similar in terms of their global structure, style, and conventions, than texts that have a different purpose and audience (i.e., texts in different genres). This implies that the writing task does not only require content knowledge concerning the topic to write on, but also knowledge on how to conform to certain standard practices within a particular genre, in order to fulfill its communicative purpose (Bhatia, 1993). Crowhurst and Piche (1979) showed that task variables such as intended audience and mode of discourse indeed affect writing products. Their study demonstrated that there is more syntactic complexity in argument than in narration or description at both Grade 6 and Grade 10 and even more when arguments were intended for a teacher than for a friend.

When writing products differ across genres, the process of writing may be different as well (Beauvais, Olive, & Passerault, 2011). Writing assessments consisting of highly homogeneous tasks (i.e., sharing the same communicative purpose) tend to underrepresent the construct of writing. Quellmalz, Capell, and Chou (1982) claimed that writing tasks in different genres (e.g., narrative versus expository writing) tap into different cognitive processes. This claim is underpinned by Reed, Burton, and Kelly (1985) who demonstrated through the use of a secondary-task procedure that genre affects cognitive capacity during writing. Proficient writers appeared to be most cognitively engaged when writing persuasive essays and least engaged when writing descriptions. Less proficient writers, on the other hand, were most engaged when writing descriptions, but least engaged when writing narratives. Beauvais et al. (2011) further showed that students adapt their writing strategy according to the genre in which they are writing. However, as was the case in earlier research, students in both studies only wrote one task in each genre. Hence, it is still unclear whether the effects are related to genre or topic. Van Weijen (2009), for instance, showed that the process of writing also differs between tasks in the same discourse mode, in this case argumentative writing.

Differences in writing products and processes suggest that performing well on one genre does not necessarily predict performance in other genres. In order to make valid inferences on writing proficiency in general, writing assessments should include multiple tasks in multiple genres. However, this will probably negatively affect the generalizability of text scores to writing proficiency, as scores based on similar tasks are more likely to correlate and, thus, to generalize. Owing to the confounded effect of topic and genre in earlier research (Coffman, 1966; Reed et al., 1985; Veal & Tillman, 1971), it is still unclear whether generalization is warranted across genres, and hence, to generalize to writing proficiency in general. Therefore, the central question in the current study is whether genre has an effect on the generalizability of text quality scores.

In order to answer this question, text quality scores of different kinds of texts written by students at the end of primary education will be analyzed. These written texts constitute four different genres, differentiated by the rhetorical purpose of the writing (narrative versus argumentative writing) and audience specification. In order to disentangle genre from topic effects, students wrote multiple tasks within the same genre. Based on the magnitude of the variance components of persons, raters, tasks, and genre, it will further be estimated how many texts, in different genres, have to be written, and how many ratings of each text are necessary in order to make inferences of differences in text quality to differences in writing proficiency.

Method

Participants

Written texts were obtained from 67 11-year-old and 12-year-old students in their final year of primary education (i.e., US Grade 6). These students were randomly selected from three different primary schools in the Netherlands. The participants were part of a larger research project on the effect of a digital writing program on students' writing performance (Pullens, 2012). The participants in the present study constituted the control group in this project (36% of all 186 participants), which received no explicit writing instructions and did not show a significant improvement of text quality scores during the period of study (Pullens, 2012).

Material and procedure

In total, students completed 12 writing tasks at three different moments during the study. At each moment, they received four paper-and-pencil writing tasks in four different genres. See Table 1 for an overview of the different tasks in four different genres, classified according to their intended purpose and audience. Narrative and argumentative writing tasks were included, as these are the two most prominent genres in both the Dutch national writing curriculum as well as the writing standards for primary education (Meijerink, 2008).

Multiple tasks were collected per genre to disentangle genre effects from topic effects. Tasks within a genre were similar in terms of audience and purpose of the texts, and differed only in topic. For instance, in argumentative writing for a specified audience,

Table 1. Different genres used in the present study, classified according to purpose and audience of the writing task.

Purpose	Audience	
	Specified reader	Unspecified reader
Argumentative writing	Persuasive letters for a fictional company	Argumentative essays to prepare oneself for a discussion
Narrative writing	Adventure stories for readers of a school newspaper	Personal stories

students were asked to write three persuasive formal letters to fictional companies about promotional campaigns. One of these letters was to a supermarket about the collection of toys, the second letter was to a petrol station about the collection of tickets to a musical, and the third letter was to a chocolate company about the collection of a music CD. See Appendix A for an overview of the specific tasks used in the present study. The format and content of persuasive letters and adventure stories in the present research were quite fixed, whereas stories about personal experiences and argumentative essays were free writing assignments.

Rating procedure

To control for the effect of handwriting on text quality ratings (McColly, 1970), the handwritten texts were retyped. Guidelines prescribed that typed texts should resemble the handwritten texts precisely, that is, all errors concerning spelling, grammar, interpunction, or capitals had to be copied, as well as all modifications made by the student. Furthermore, students' names were hidden and replaced by unique codes to preserve anonymity.

For efficiency reasons, there was a design of overlapping rater teams (Van den Bergh & Eiting, 1989). Through this rating procedure, each essay was rated by a jury of three raters, without the need for having raters assess all the essays. Rater juries were selected from a total of 32 student teachers. The persuasive letters were also rated by juries of three experienced raters, who were randomly selected out of 17 teachers with at least five years of experience in the upper grades of primary education. By investigating whether ratings from inexperienced raters differ from experienced raters it was possible to analyze the effect of rater expertise on score generalizability and to determine whether results should be corrected for the raters' background.

Texts were holistically rated using benchmarks that represent the (approximate) average text quality for the writing task in question. Earlier research (Blok, 1986; Schoonen, 2005; Tillema, Van den Bergh, Rijlaarsdam, & Sanders, 2012) has shown that this rating procedure considerably increased rater reliability. After elaborate inspection of the sample of written texts, a benchmark for each writing task was selected by experienced raters. It was assured that the selected benchmarks did not contain too many grammar and spelling errors. This was important because otherwise raters' attention could be drawn away from the content and structure of the text towards mechanics. For each benchmark was explained what was average about the text according to specified criteria, such as content,

Table 2. Reliabilities of individual raters ($N = 32$) and jury raters ($N = 3$; 17 jurors) per writing task.

Genre	Task	Reliabilities of individual raters (ρ, SD)	Reliabilities of juries of three raters (ρ, SD)
Persuasive letters	1 Collecting toys	.65 (.30)	.83 (.12)
	2 Collecting musical tickets	.64 (.25)	.83 (.09)
	3 Collecting music CD	.62 (.24)	.82 (.10)
Argumentative essays	1 Candy prohibition	.50 (.31)	.74 (.11)
	2 Smoking ban	.53 (.26)	.76 (.09)
	3 Telling tales	.58 (.24)	.79 (.09)
Adventure stories	1 Sports field	.58 (.33)	.80 (.06)
	2 Forest fire	.57 (.24)	.78 (.11)
	3 Poison	.68 (.19)	.86 (.07)
Personal stories	1 Being frightened	.54 (.30)	.77 (.07)
	2 Being caught	.49 (.37)	.73 (.10)
	3 Home alone	.61 (.22)	.82 (.06)

structure, style, conventions, and mechanics. Raters had to compare the student texts to the benchmarks, and they had to assign a score to the texts, indicating the extent to which they thought the texts were better or worse than the benchmark. As ratings are at best interval scaled (Suppes & Zinnes, 1963), each benchmark was given an arbitrary score of 100. Thus, if a rater thought a text was twice as good as the benchmark, the text was awarded a score of 200. And vice versa, a text that was considered half as good as the benchmark, according to the raters, received a score of 50. To become familiarized with the procedure, raters received a short training procedure before actually marking the essays. During this training they practiced rating several writing examples of varying quality, and discussed their judgments.

Application of a multi-group LISREL model on the covariance matrix between raters (Van den Bergh & Eiting, 1989) provides estimates of the reliability of ratings of each rater in relation to all other raters of that task (see Table 2). As expected, the individual reliabilities are not high (cf. Godshalk et al., 1966; Huot, 1990b; McColly, 1970). As each essay was rated by three raters, the reliability of each jury can also be estimated (see Table 2). These jury reliabilities are quite satisfactory, and the variance between each jury's rating of essays of the same task can be considered relatively low.

Design and analysis

Since the aim of the current study is to differentiate between skilled and unskilled writers, students were the object of measurement. The other random facets included in the research design were as follows: (1) genre (g), fully crossed with persons (p); (2) tasks (t), nested within genres; and (3) raters (r), nested within tasks. This resulted in a partially nested, three-facet univariate design (p * (r:t:g)).

In total there were 2207 text scores available. Due to absence, not all students were able to complete all 12 tasks. Moreover, some texts were only rated two instead of three times. As a consequence, 205 of the total 2412 (8.5%) observations were missing, resulting in an unbalanced design.

To estimate the generalizability of writing scores, variance components were calculated for each of the seven sources of variance possible in the research design: person, genre, person by genre, task within genre, person by task within genre, raters who rated tasks within different genres, and random error. The variance components were estimated by means of the restricted maximum likelihood (REML) approach in SPSS. REML was used in order to obtain best linear unbiased estimates in unbalanced designs (Searle, 1987).

Several G-studies were performed in order to analyze the effect of genre on the generalizability of writing scores. First, to estimate the relative influence of genre, a G-study was performed in which genre was considered to be a random facet in the measurement. Second, to determine whether the stability of students' writing performance differed from genre to genre, a G-study was performed for each condition of the fixed facet genre. The expectation is that, when limiting the universe of generalization to a given genre, variance between persons will at least be as high as, or higher than, the variance for writing in general. Third, we performed an extra G-study on the ratings of persuasive letters by experienced teachers. The generalizability of these ratings was compared with the generalizability of ratings given by student teachers, to determine whether rater experience affects score generalizability.

In D-studies we approximated how many tasks and raters were needed to attain a reliable judgment about writing proficiency, both within and across genres. Estimations of variance components were used to compute generalizability coefficients for relative decisions and dependability indices for absolute decisions according to varying numbers of raters and tasks. Generalizability coefficients differ from dependability indices in what is considered to be measurement error (Cronbach et al., 1972). Calculations of generalizability coefficients were based on the ratio of person variance to measurement error influencing only the *ranking* of persons. That is, all interaction effects with persons, specifically, interaction effects of person-by-genre, person-by-task and random error including the three-way interaction of person-by-rater-within-tasks. Dependability indices were calculated by the ratio of person variance to all sources of error including main effects of genre, tasks and raters.

Results

Genre as a random facet in the measurement

The central question concerns the generalizability of text quality scores: is generalization over genres and tasks within genres warranted if students only write one text in one genre? To estimate the generalizability, the observed score variance is decomposed into seven variance components: the variance due to persons, genres, tasks within genre, raters who rated tasks within different genres, their interactions and random error. In Table 3, the percentages of variance associated with each of these components are summarized.

Results show that the person variance, the component of interest, only accounts for 10% of the variance in text scores. Hence, the correlation between individual text quality

Table 3. Variance decomposition, in percentages, for persons, tasks and raters, separated by genre.

Source (facet)	Percentage of variance
Person (p)	9.98
Genre (g)	11.42
Person by genre (pg)	4.01
Task within genre (t:g)	1.71
Person by task within genre (p(t:g))	19.13
Rater within task and within genre (r:t:g)	18.05
Person by rater within tasks, within genre, and error (p(r:t:g), e)	35.71

scores on random written texts, rated by one randomly selected rater, is low, only .32 on average. Thus, text quality scores largely (for 90%) depend on facets that are not directly related to individual writing proficiency.

First of all, genre appears to be an important facet in the design. The main effect of genre accounts for 11% of the variance. Thus, average writing scores differ between genres, indicating that genres differ in difficulty; scores of text quality are slightly more similar within genres than between them. If the facet genre is not included in the analyses, the proportion of variance related to differences between persons will be overestimated. This is especially the case for decisions about writing proficiency based on absolute levels of text scores, because in these instances, genre is considered to be part of the measurement error. For instance, the observed text score of a student writing in a relative easy genre will be higher than when the same student writes a text in a more difficult genre. Decisions based on absolute writing scores are therefore affected by genre, indicating that one should write texts in more genres in order to be able to generalize to writing proficiency.

When decisions only concern the ranking of persons (i.e., relative decisions), genre is of considerably less influence: only 4% of the variance is due to the interaction of person by genre. This shows that, although the quality of a text is affected by the difficulty of a genre, better writers still outperform worse writers. This conclusion does not hold for tasks *within* a genre: the ranking of persons varies widely over tasks within a specific genre, as indicated by a variance component of almost 20%.

Besides the effects of genre and topic, 18% of the variance in text quality scores is explained by the interaction of rater within tasks within genre, and more than one third (35%) by random error, including the interaction of persons by rater within tasks and within genre. This residual variance is difficult to interpret, because of confounding variables in the design. It does, however, indicate that scores within and between persons fluctuate enormously, owing to differences in raters and how raters rate different tasks from different genres.

Decisions about writing proficiency across genre

In order to make an approximation about how many writing tasks and raters are needed to attain a reliable absolute judgment about writing proficiency in general, the

dependability index was estimated in an absolute D-study. The dependability index is the ratio of person variance to all the variance in text scores, including unwanted variance related to all measurement facets (measurement error). The D-study showed that, in order to reach the desired level of dependability of at least .70, students should write at least four different texts in six different genres, that is, a total of 24 texts. These texts should be rated by at least three different raters. For relative decisions about writing performance, comparable levels of generalizability are attained with only 12 texts based on three different writing tasks in four different genres, rated by only two raters. Relative decisions are, however, only valid decisions when the writing assessment is used for norm-referenced testing in which the goal is to determine the relative performance of students in comparison. Relative decisions do not provide information about whether students meet a fixed standard of writing, that is, criterion-referenced testing.

All in all, the results indicate that the proportion of variance in writing scores that is explained by individual differences is rather small compared to the large effects of measurement aspects, such as rater, genre, interactions between person and genre, person and task and random error including the three-way interaction of person by rater within task.

Genre as a fixed facet in the measurement

To see whether the stability of students' writing performance differed from genre to genre, a G-study was performed for each condition of the fixed facet genre: persuasive letters, argumentative essays, adventure stories and personal stories. Moreover, for writing scores of persuasive letters, an extra G-study was performed to the ratings of experienced teachers in order to compare these ratings with ratings from student teachers. Table 4 summarizes the proportions of variance components (in percentages) for these G-studies.

Again, persons only accounted for a relatively small part of the variance in text quality scores (12–24%). As expected, this varies between genres; tasks in which students have to write about personal experiences show relatively more variance between students (24%) than tasks in the other genres (persuasive letters, argumentative essays or adventure stories, 12–18%).

As expected, the results do not show a main effect of tasks. The rating procedure was equal throughout all genres: every text had to be compared to a benchmark text of 100 points. For all tasks, the mean of the text quality scores across students was therefore approximately 100 points. Although scores did not vary systematically across tasks, persons performed differently on different tasks – the interaction of person by task ranged from 17% for personal stories to almost 30% for persuasive letters.

Furthermore, in line with the results presented in Table 4, there is a large interaction effect of rater within tasks. Judgments of raters vary from 14% for personal stories to 24% for persuasive letters and argumentative essays. This indicates that differences between raters' judgments depend on the genre of the rated texts. Specifically, raters were most familiar with the characteristics of a good personal story. To see whether consistencies in ratings were affected by rater experience, experienced teachers' ratings for persuasive letters were compared to student teachers' ratings. As expected, ratings within tasks were somewhat more consistent for experienced teachers (accounting for 18% of

Table 4. Percentage of variance due to measurement facets for different genres and rater panels differing in experience.

	Teachers	Students			
	Persuasive letters	Persuasive letters	Argumentative essays	Adventure stories	Personal stories
Person (p)	17.50	14.66	12.40	12.26	24.45
Task (t)	0.27	2.33	0	3.09	1.12
Person by task (p*t)	29.60	22.63	18.35	25.92	17.39
Rater within task (r:t)	17.89	23.94	24.32	18.82	13.60
Error (p*(r:t), e)	34.73	36.45	44.94	39.90	43.44

the variance), than for student teachers (accounting for 24% of the variance). More important, however, is the impact on generalizability, indicating that there is hardly any difference between experienced and student teachers' scores – differences in text quality are only slightly related to differences in individual writing proficiency.

Decisions about writing proficiency within specified genres

Although the stability of writing proficiency depends on the type of texts to be written, for every genre there are still other sources than students' writing proficiency at stake. The implication was that multiple tasks and multiple raters are necessary in order to generalize text quality scores to writing proficiency in a specific genre. Relative and absolute D-studies were performed to approximate the number of tasks and raters for reliable measurement of genre-specific writing.

In Figure 1 generalizability coefficients are plotted for multiple tasks (*x*-axis) and multiple raters (lines). The figures show that in order to generalize (.70 or higher) beyond given tasks or raters, one needs at least five tasks and five raters for persuasive letters, argumentative essays or adventure stories. In contrast, only three tasks and three raters are necessary for writing personal stories. Dependability indices for absolute decisions are lower in all four genres. Specifically, for five tasks and five raters, dependability indices are .66 for persuasive letters, .66 for argumentative essays, .60 for adventure stories and .80 for personal stories.

Discussion

In the current study, aspects of the measurement of writing were disentangled in order to investigate the validity of inferences made on the basis of writing performance and to define implications for the assessment of writing. By including genre as a facet in the measurement, we obtained writing scores of 12 texts in four different genres for each student. Results indicate that across raters, tasks and genres, only 10% of the variance in writing scores is related to individual writing skill. Thus, when tasks are considered as a random selection of the universe of all possible tasks, it is quite hard to draw generalizable conclusions about writing proficiency beyond the given rater and task. More specifically,

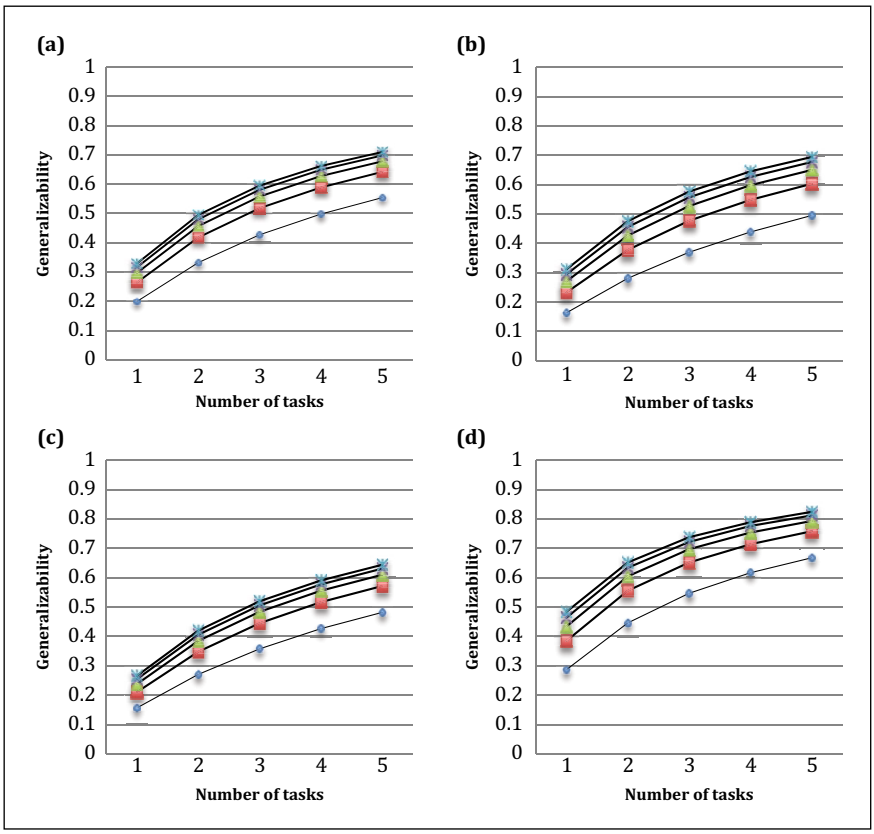

Figure 1. Estimated generalizability of writing scores for relative decisions, with varying number of tasks and raters within four different genres: (a) persuasive letters; (b) argumentative essays; (c) adventure stories; (d) personal stories. Lines within a graph represent number of raters, ranging from one rater (lowest line) to five raters (upper line).

for valid and reliable inferences on relative writing performance, students should at least write three texts in each of four genres, rated by at least two raters. Even more tasks and raters are necessary for absolute decisions about writing proficiency. If it is not feasible to include different tasks in multiple genres in the writing assessment, writing scores may also be obtained by fewer but more similar kinds of tasks, that is, tasks within one genre. Absolute inferences should in this case be limited to genre-specific writing, as the results of the present study show that generalization across genre is not warranted when writing scores are obtained from texts within the same genre.

However, it should be noted that only when information is available about task compatibility or task equivalence, it is allowed to determine absolute performance levels in writing. As ratings of writing quality in writing education or writing research are generally measured on interval level, at best, differences between tasks are quite arbitrary. In this case it is not beneficial to add more tasks and raters to the assessment of writing.

Our findings emphasize the effect of genre on the generalizability of writing scores. This is in line with earlier research, showing that writing performance differs substantially between different kinds of writing tasks (Coffman, 1966; Crowhurst & Piche, 1979; Moss, Cole, & Khampalikit, 1982; Quellmalz et al., 1982; Reed et al., 1985; Schoonen, 2005; Van den Bergh et al., 2012; Veal & Tillman, 1971). However, until now, analyses of task effects were almost always based on single tasks within multiple genres (Coffman, 1966; Reed et al., 1985; Veal & Tillman, 1971) or on multiple tasks in one genre (Van den Bergh et al., 2012, Van Weijen, 2009). Hence, topic and genre effects were confounded and, as a result, inferences that could be drawn about systematic differences within and across genres were limited. The current research extends this knowledge by untangling the effects of topic and genre by including multiple tasks in multiple genres in the measurement. The results show that genre has an effect above and beyond specific task effects. This implies that if the facet genre is not included in the analyses, the proportion of variance related to differences between persons will be overestimated (see, e.g., the large person variance in Van den Bergh et al., 2012).

The results in the current study also show that the generalizability of writing scores differs from genre to genre. Presumably, for young students, it is easier to generalize to personal writing (only three texts rated by three raters are necessary) than to persuasive writing (at least five texts rated by five raters are necessary). A possible explanation for this effect is genre knowledge. When individuals are familiar with certain communicative goals, they internalize genre-specific conventions in order to reach these goals in a standardized, and thus efficient, way. As a result, writing within specific communicative events is expected to be more stable for expert members of this event (Bhatia, 1993; Swales, 1990). As young students in primary education, the object of measurement in the present study, routinely communicate about personal experiences in school, it is assumed that they have well-developed schemata for personal writing. Their genre knowledge importantly influences choices for content, structure and rhetorical style in writing. Therefore, compared to other genres, texts about personal experiences will be better comparable, and, better generalizable. Vice versa, for genres that are not practiced regularly in school, students are likely to approach each writing task as a new one, thereby limiting generalizability across tasks.

The intended audience, as specified by the writing task, may also explain the previously mentioned genre differences. Although results related to audience specification were somewhat mixed, writing performance appeared to be more stable for personal writing than for writing for specific readers. This could indicate that, at least for young students, writing about oneself is easier than writing for someone else. Bereiter and Scardamalia (1987) indeed argued that young students experience difficulties in transforming ideas and knowledge to reach a specific audience. Contrary to this theorization are the results of the quality of argumentative essays, which varied as much as the other genres. For these essays, students were asked to write down arguments in preparation for a class discussion, with the student himself as the intended audience. However, as their teacher was the leader of the subsequent class discussion, students might have written argumentative essays with the teacher as intended audience. This could have affected variability in students' writing performance, because Grade 6 students experience more difficulties when writing arguments for their teacher than writing for a friend, resulting

in higher score variance (Crowhurst & Piche, 1979). Research on writing assessment is not conclusive about the relationship between components of the writing task, such as communicative purpose and intended audience, and writing quality (Huot, 1990b). It is therefore necessary to study these effects more systematically in further research.

Further, results show that writing performance differs *within* genre. Even when genre is included in the analysis, large variance between text quality ratings is observed, due to the interaction effect of person-by-task-within-genre. Because the estimated task effects are contaminated by effects of genre, it is not clear whether task effects are due to topic knowledge, task familiarity or due to the interaction of task-by-genre. However, it seems almost impossible to differentiate between topic and discourse mode, because choice of topic may constrain choice of public or communicative goal, and vice versa.

As expected, raters were not consistent in their judgment of text quality. This study confirms findings from previous research (Godshalk et al., 1966; Huot, 1990a) that text quality is reflected best by writing scores based on judgments of multiple raters. In line with earlier research (Gebril, 2009; Lee & Kantor, 2007), however, the effects of rater, including rater-by-task interaction, appear to be smaller than the effects of task, including genre and topic effects. This was true in all observed genres. Overall, it was estimated that the effects of task, genre and topic effects included, accounted for twice as much of the variance in writing scores as those of raters. Although the estimated residual error variance includes both rater and task effects, it seems that, ceteris paribus, it is more efficient to increase the number of tasks in writing assessment, than the number of raters.

Previous research indicated that rater experience might be an explanation for rater variability (Barkaoui, 2007; Schoonen, 2012). However, in this research, there was no effect of rater experience, at least not for the judgments of persuasive letters. Ratings of experienced teachers are more precise and consistent over tasks, given the smaller interaction effect of rater by task, but they are not superior to judgments of student teachers, as ratings of experienced teachers show larger interaction effects of person by task.

Hence, the present study shows that writing performance largely depends on the writing task. This raises the important question of how to interpret this finding. In general, and in line with generalizability theory, writing proficiency is considered to be a relatively constant disposition of a person. Writers are assumed to perform in a more or less consistent way across tasks. In terms of generalizability coefficients, this means that writing proficiency comprises only shared variance between tasks; all other score variance is considered to be measurement error. It can, however, be questioned whether task effects or the interaction of person-by-task should be regarded as measurement error. For instance, there are researchers who consider specific task effects as part of writing proficiency (Chalhoub-Deville, 2003; Read & Chapelle, 2001; Verheyden, 2011). This interactionalist view of language performance assumes that writing proficiency interacts with its context, implicating that changes in the writing task (i.e., context) will lead to changes in text quality (writing performance). This, however, has significant implications for the generalizability of the inferences made, and thus of the definition of the construct of writing proficiency (for similar reasoning, see Schoonen, 2012).

Even when writing is specified to narrower domains, writing performance is likely to fluctuate across tasks. Large task effects could, at least partly, be reduced by proper education in writing. Earlier research has shown that students do not regularly practice

writing in class and that, when they do write, they hardly receive any feedback on their writing process or product (Henkens, 2010). Students therefore lack an effective approach to writing that would lead to a more consistent writing performance of higher quality. Hence, students' writing could be improved by learning effective strategies for transferring general writing principles to novel writing tasks.

Parkes (2001) has already hypothesized that task variance in performance assessment may be reduced by facilitating the transfer of knowledge across performance situations. According to his review of literature on transfer issues, effective transfer largely depends on a subtle balance between general knowledge and specific situations. He proposes three broad solutions for the transfer problem. First, students should get a good understanding of the general approaches to cognitive problem solving. In the context of writing, this means that students should know the characteristics of a good text, and how content, structure and style relate to effective writing. Second, students should form general schemata of applying general principles to specific tasks. In order to learn how to use these schemata, concrete examples of tasks should be provided. Moreover, writing strategies will help students to produce texts of a more or less consistent quality. Third, tasks should be well defined, for instance, by making the goals of the task explicit. Clear tasks promote students to see the analogy between tasks and to match the general model of earlier experiences to new situations. A writing task, therefore, should contain explicit information about the communicative goal, the topic and the intended audience. Future research should empirically test whether these solutions indeed affect the ability to transfer general writing knowledge to specific writing tasks, and thereby, reduce task variance.

It is very likely that some other features in this study have affected the outcomes. For instance, ratings of text quality were task dependent, as benchmarks differed between tasks. This does not necessarily affect the comparability of text quality scores *between* tasks, as text quality was measured at best on interval level. As a consequence, task variance may be underestimated, making it difficult to interpret absolute decisions. Moreover, it is possible that raters used different criteria for rating text quality per task, which could result in artificially high estimates of the interaction of person by tasks. After all, by applying different criteria to the tasks, the rank order of students might differ between tasks; good writing in one task does not necessarily imply good writing in another task. Nevertheless, jury reliabilities appeared to be acceptable for all writing tasks, suggesting that raters marked the essays in more or less the same way. It is thus more likely that rater variance contributes to random noise related to interactions between rater and text. Whereas rating criteria may vary for texts in different genres (e.g., persuasive writing is rather different from storytelling), criteria for genre-specific writing should be more or less alike. Further research should therefore use rating procedures that support raters in a more task-independent way, at least for rating texts that are similar in terms of their communicative purpose and audience. Recent research has already suggested that benchmarks can be used for different tasks within the same genre (Tillema, 2012). This is in line with the finding that benchmarks promote raters to judge texts as a whole (Schoonen, 2005).

In writing research as well as in writing education, writing proficiency is still quite often assessed with a single writing task rated by one rater. The current study shows,

however, that decisions regarding writing proficiency based on one written text are not very reliable. Neither are decisions on multiple, but highly similar, texts. Because the ability to write differs from genre to genre, generalizable inferences are not appropriate. In order to draw conclusions about writing in general, writing assessment should rather include multiple tasks in multiple genres rated by multiple raters. If it is not possible to include multiple texts in different genres, for instance, because of time or money constraints, decisions should be limited to genre-specific writing.

Funding

This research has been supported by the Netherlands Organization for Scientific Research (NWO), grant 411-11-859 to Huub van den Bergh. We are grateful to Theo Pullens for providing the data.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238–257.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. doi:10.1016/j.asw.2007.07.001

Beauvais, C., Olive, T., & Passerault, J. M. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology*, *103*(2), 415–428. doi:10.1037/a0022545

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Mahwah, NJ: Lawrence Erlbaum.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Blok, H. (1986). Essay rating by the comparison method. *Tijdschrift voor Onderwijsresearch*, *11*, 169–176.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369–383.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement*, *3*(2), 151–156. doi:10.1111/j.1745–3984.1966.tb00872.x

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: John Wiley.

Crowhurst, M., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English*, *13*(2), 101–109.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*(4), 507–531.

Godshalk, F. I., Coffman, W. E., & Swineford, F. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Henkens, L. S. J.M. (2010). *Het onderwijs in het schrijven van teksten* (pp. 1–58). Utrecht: Inspectie van het Onderwijs.

Hoetker, J. (1982). Essay examination topics and students' writing. *College Composition and Communication*, *33*(4), 377–392. doi:10.2307/357949

Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement*, *69*(6), 887–912.

Huot, B. (1990a). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201–213. doi:10.2307/358160

Huot, B. (1990b). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237–263.

Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing*, *7*(4), 353–385.

McColly, W. (1970). What does educational research say about the judging of writing ability? *Journal of Educational Research*, *64*(4), 147–156.

Meijerink, H. (2008). *Over de drempels met taal en rekenen* (pp. 1–56). Enschede: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. *Journal of Educational Measurement*, *19*(1), 37–47.

Parkes, J. (2001). The role of transfer in the variability of performance assessment scores. *Educational Assessment*, *7*(2), 143–164.

Pullens, T. J. M. (2012). *Bij wijze van schrijven* (dissertation). Utrecht University.

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*(4), 241–258. doi:10.1111/j.1745–3984.1982.tb00131.x

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32.

Reed, W. M., Burton, J. K., & Kelly, P. P. (1985). The effects of writing ability and mode of discourse on cognitive capacity engagement. *Research in the Teaching of English*, *19*(3), 283–297.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27, pp. 1–22). Leiden: Brill. doi:10.1108/S1572–6304(2012)0000027004

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*(2), 157–184.

Searle, S. R. (1987). *Linear models for unbalanced data*. New York: John Wiley.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Suppes, P., & Zinnes, J.L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology* (vol. 1, pp. 39–74). New York: John Wiley.

Swales, J. (1990). *Genre analysis*. Cambridge: Cambridge University Press.

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes*. Utrecht: LOT Dissertation Series.

Tillema, M., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2012). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*, *30*(1), 1–27. doi:10.1177/0265532212442647

Van den Bergh, H., & Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement*, *26*(1), 29–40.

Van den Bergh, H., Maeyer, S. de, Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. Van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (vol. 27, pp. 23–32). Leiden: Brill.

Van Weijen, D. (2009). *Writing processes, text quality, and task effects; empirical studies in first and second language writing*. Utrecht: LOT Dissertation Series.
Veal, L. R., & Tillman, M. (1971). Mode of discourse variation in the evaluation of children's writing. *Research in the Teaching of English*, *5*(1), 37–45.
Verheyden, L. (2011). *Achter de lijn. Vier empirische studies over ontluikende stelvaardigheid* (dissertation). Katholieke Universiteit Leuven.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150–173. doi:10.1016/j.asw.2011.12.001

Appendix A

Table A1. Writing tasks that are used in this study: three topics in four different genres, categorized according to purpose of writing and audience specification.

Purpose	Audience	
	Specified reader	Unspecified reader
Argumentative writing	Persuasive letters for a fictional company Task 1: Collection of toys at a supermarket Task 2: Collection of stamps at a petrol station for earning musical tickets Task 3: Collection of points on wraps of chocolate bars to earn a music CD	Argumentative essays to prepare oneself for a class discussion Task 1: Pros and cons of a candy prohibition for children Task 2: Pros and cons of a smoking ban Task 3: Pros and cons of telling tales about somebody
Narrative writing	Adventure stories for readers of a school newspaper Task 1: Adventure on a sports field Task 2: Adventure about a forest fire Task 3: Adventure about poison	Personal stories Task 1: Personal experience about being frightened by something Task 2: Personal experience about being caught for something Task 3: Personal experience about being home alone