Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Assessing Writing 9 (2004) 105–121
ASSESSING WRITING

Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system

Gavin T.L. Brown[a,*], Kath Glasswell[b], Don Harland[c]

[a] *School of Education, Private Bag 92019, University of Auckland, Auckland, New Zealand*
[b] *University of Illinois, Chicago, USA*
[c] *Howick College, Auckland, New Zealand*

Available online 25 July 2004

Abstract

Accuracy in the scoring of writing is critical if standardized tasks are to be used in a national assessment scheme. Three approaches to establishing accuracy (i.e., consensus, consistency, and measurement) exist and commonly large-scale assessment programs of primary school writing demonstrate adjacent agreement consensus rates of between 80% and 100%, and consistency and measurement coefficients ranging between .70 and .80, .60 and .80, respectively. A New Zealand educational assessment project has developed a set of writing assessment rubrics that contain curriculum based rating scales for six purposes of writing each with its own bank of writing prompts for use by classroom teachers. Standardization of the rating scales and prompts was conducted with representative samples of students, aged 10–13. This article describes two studies that established the validity of the scoring system for use in New Zealand classrooms. Adjacent agreement consensus fell between 70% and 90%, while consistency and measurement correlations fell in the range .70–.80 in both studies. This consistency with international standards was sufficiently robust to provide confidence in the underlying norms provided by the asTTle assessment tool. Relatively low levels of training were required by teachers to reach this degree of accuracy. The accuracy of scoring gives government and teachers confidence in the validity of the project's rating scales and suggests that classroom teachers will be able to generate accurate scores upon which instructional decisions can be based.

Keywords: Reliability; Validity; New Zealand; Large-scale assessment; Assessment tool

* Corresponding author. Tel.: +64 9 3737 599x83042; fax: +64 9 308 2355.
E-mail address: gt.brown@auckland.ac.uz (G.T.L. Brown).

1075-2935/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.asw.2004.07.001

1. Introduction

The accurate scoring of writing is a ncessary requirement of any large-scale system of national assessment and yet this represents one of the more serious challenges in the design of such national assessments. Various mechanisms have been demonstrated as effective in improving the accuracy of scoring, including: use of explicit scoring rubrics (Gearhart, Herman, Baker, & Whittaker, 1992; Huot, 1990; Linn & Gronlund, 2000; Popham, 2000); use of a reasonably long scoring scale (Coffman, 1971); use of augmentation of holistic grades (e.g., adding plus or minus) (Penny, Johnson, & Gordon, 2000); cross-checking or moderation of marking (Gronlund & Linn, 1990); systematic scoring processes (Airasian, 1997; McMillan, 2001); training of markers (AERA/APA/NCME, 1999; Breland, Bridgeman, & Fowles, 1999), the use of expert markers (Penny et al., 2000); and the use of panels (AERA/APA/NCME, 1999; Breland et al., 1999). The purpose of this paper is to overview the literature related to the demonstration of reliability in the scoring of writing, describe a New Zealand writing assessment program, and report levels of accuracy obtained in two studies which used the writing assessment program.

2. Determining accuracy in the scoring of writing

Three main approaches to determining the accuracy and consistency of scoring exist (i.e., consensus estimates, consistency estimates, and measurement estimates) and it has been argued that establishing inter-rater accuracy requires all three approaches. Consensus estimates indicate the degree to which markers give the same scores, consistency estimates indicate the degree to which the pattern of high and low scores is similar among markers, and measurement estimates indicate the degree to which scores can be attributed to common scoring rather than to error components (Stemler, 2004).

Consensus estimates are used when raters are trained to judge according to rating scale rubrics assumed to represent a linear continuum of progress in a construct. Popular indices of consensus agreement include percent exact agreement (i.e., what percentage of scores awarded are exactly the same) and percent adjacent agreement (i.e., what percentage of scores within plus or minus one score category of each other). Cohen's kappa statistic (Cohen, 1960) can be used to determine the degree to which consensus agreement ratings vary from the rate expected by chance, with values greater than .60 indicating substantial non-chance agreement. Obviously, high scores are much easier to achieve, using the adjacent agreement approach than the exact agreement approach and high scores are easily achieved when the rating scale has few categories (i.e., up to four points) (Stemler, 2004). Nevertheless, exact agreement levels of 70% or more are considered indicative of reliable scoring (Stemler, 2004).

Consistency coefficients (such as Pearson for pairs of raters or Cronbach's alpha for multiple judges) indicate that there is a noticeable pattern in the distribution

of scores across a pool of raters rather than whether raters have given the same scores. A high coefficient indicates that raters gave high and low scores in a similar pattern across a body of commonly rated events or performances. However, the mean scores of each rater may be quite different, because high for one judge is not necessarily the same point on the rating scale as it is for another rater. Thus, adjustments for relatively harsh or lenient raters may be needed. A further weakness of this approach to estimating accuracy is that when there is little variance among raters, for example, when they give exactly the same scores, correlation coefficients will be deflated and may lead to the false conclusion that consistency was poor.

Measurement estimates apportion variance in assigned scores to task, rater, error, and interaction components and thus provides the most robust indicator of degree of agreement attributable to the similarity of raters' scores (Shavelson & Webb, 1991). Such estimates (e.g., phi index of dependability from generalisability theory) can establish the degree to which a score represents a 'true' score taking into account the effect of all judges' severity and internal consistency and any interaction of task, rater, and error. Coefficient values exceeding .80 are accepted as robust indication that judges are rating a common construct (Shavelson & Webb, 1991).

This section of the paper reviews various published accuracy statistics related to the scoring of writing. Special attention has been paid to the assessment of writing in primary and secondary school contexts as they are critical to the empirical work described later. In 1990s, three large-scale research programs in the scoring of primary school students' writing have been launched with extensive analysis and publication of results (i.e., The New Standards Project, CRESST's development of new narrative-specific scoring rubrics, and Vermont's portfolio assessment scheme).

The New Standards Project is an effort to create a national examination system with strong reliance on performance tasks closely tied to the curriculum in use in each state's educational setting, but permitting comparison to a common national performance standard. The project reported substantial variation in state-by-state mean scores for writing samples at each of three grade levels when scored against the writing rubrics valid in each state; for example, at the middle school level the same scripts scored in three or more states obtained mean scores separated by .1 to .2 standardized *z* scores (Linn, 1991; Linn, Kiplinger, Chapman, & LeMahieu, 1992). However, median consistency coefficients ranging .73 to .81 were found in the New Standards Project inter-state study of writing scoring in elementary, middle school, and high school grades even though each state used its own scoring rubrics. A pilot study of the New Standards Project on-demand writing tasks found an average consistency coefficient of only .54 across three tasks scored holistically, using a seven-point scale (Resnick, Resnick, & De Stefano, 1993). It was also found that exact agreement percentages ranged between 40% and 49% and adjacent scoring percentages of between 86% and 88%, using a seven-point rubric for three writing tasks scored by 114 teachers.

The Center for Research on Evaluation, Standards, and Student Testing (CRESST) has been developing and testing the reliability and validity of novel

writing rubrics for use in assessing primary school children's writing. Comparing analytic and holistic six-point scoring rubrics of Grade 3 writing, Gearhart et al. (1992) found exact agreement percentages ranging from 15% to 97%, and adjacent agreement percentages ranging from 80% to 100% and an average Pearson consistency coefficient of .70 was found. Related work comparing portfolio scoring by three raters with classroom writing found consistency correlations around .60 to .62 between scores for portfolio portfolios and in-class narrative and summary writing at Grade 3 (Herman, Gearhart, & Baker, 1993). In contrast, later work in the same program of research found that six-point narrative-specific and general analytic scoring rubrics produced consensus exact agreement results of 39%–46% and 28%–37%, respectively, and adjacent agreement results of 95%–97% and 92%–94%, respectively, with average consistency correlations ranging from .48 to .68 (Gearhart, Herman, Novak, Wolf, & Abedi, 1994). The same study reported measurement coefficients ranging from .47 to .68 for two different six-point scoring rubrics; results very similar to consistency coefficients reported above. Extending this work further, a small-scale study of paired scoring using a holistic six-point rubric for overall narrative effectiveness produced a measurement approach dependability coefficient of .75 (Novak, Herman, & Gearhart, 1996).

Early in the 1990s, the state of Vermont introduced a state-wide primary school performance-based portfolio assessment program in literacy and numeracy and has monitored the robustness of the innovation. Koretz, McCaffrey, Klein, Bell, and Stecher (1993) reported exact agreement percentages of only between 44% and 48% on a four-point scoring rubric for best writing pieces in a state-wide portfolio assessment scheme and found that classroom teachers and volunteer scorers in Vermont provided almost identical average grades for the same pieces of writing. In contrast, Vermont's Uniform Test of Writing, in which students respond to common prompts, reported coefficients of .75 in the fourth grade and .67 in the eighth grade while the writing component of an experimental portfolio assessment scheme found much lower average consistency coefficients ranging between .34 and .43 for all pieces in a portfolio. Despite efforts to improve this low level of consistency, the following year's Vermont portfolio assessments in writing produced consistency correlations of only .03 to .07 better (Koretz, Klein, McCaffrey, & Stecher, 1993).

Evaluation of writing samples has also increasingly been used in various portfolio assessment schemes. In Rochester, New York, for example, it was found that 20 trained raters and students' own classroom teachers had exact agreement consensus on the writing score for K-2 students' writing portfolios, using an eight-point rubric, between 63% and 72% of the time and had consistency coefficients ranging between .68 and .73 (Supovitz, MacGowan, & Slattery, 1997). District-wide portfolio assessment in Pittsburgh, Pennsylvania of Grade 6–12 writing, by 25 trained raters, using a six-point rubric, found consensus rates, using adjacent agreement of 87%–98% and consistency coefficients of between .75 and .87 (LeMahieu, Gitomer, & Eresh, 1995).

In a study of scoring Grade 5, low-stakes writing samples, using a six-point holistic rubric, Penny et al. (2000) found that increasing the marking scale by allowing for augmentation of each grade resulted in consistency coefficients around .74 to .75 and consensus exact agreement rates of between 59% and 63% for raters and experts, respectively. Furthermore, they found that the augmented scale over the six-point holistic scale generated measurement coefficients nearly .10 greater and the use of expert raters in grading the samples increased measurement coefficient from .60 to between .72 and .85. A comparative study of trained and untrained pools of raters in Louisiana, using a six-category, four-point rubric, found measurement coefficients of between .61 and .66, with little effect attributable to training the markers (Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999).

Research into the accuracy of scoring writing is vital in the context of high-stakes examinations. Significantly more data on accuracy estimates are available for assessments at the transition point into tertiary education (e.g., ETS's Advanced Placement and TOEFL examinations), while less has been found relative to primary age children. The Advanced Placement (AP) examinations in the United States are a voluntary participation program by which students can gain entrance or accelerated progress through first year university courses. Each AP examination consists of machine and human scored elements. The Educational Testing Services routinely analyze, because of the high-stakes consequences of these public examinations, the inter-rater reliability of the human scored components. Inter-rater consistency coefficients of between .76 and .85 have been typically found in the examination of language and literature subjects, whereas the consistency coefficients in the human scoring of many science and mathematics subjects ranged from .90 to .98 (College Board, 2001; Longford, 1994). Longford (1994), nevertheless, also reported that consistency coefficients in scoring AP composition had ranged from as low as .50 to .73. Lee (2001) reported that the more than 150,000 computer-based compositions tested within the TOEFL in 1999 produced, using a six-point scale and two readers per composition, 96% adjacent agreement consensus estimates and .84 alpha consistency coefficients. Breland et al. (1999) review of major tertiary entrance writing examination systems used in the United States indicated that single-essay tests even with two raters have consistency coefficients in the range .50 to .60, while those with two essays scored by two raters have consistency coefficients in the range .70 to .80.

In United Kingdom, high-stakes assessment of writing has been extensively used and a recent study reviewed consistency of scoring between 1995 and 2002 at the Key Stage 2 (Year 6, age 11) level (Green, Johnson, O'Donovan, & Sutton, 2003). The study involved scoring writing from matching samples of children from the same schools some seven years apart. It was found that consistency correlations between the students' classroom teacher rating of writing and that of the central authority ranged in 1995 between .59 and .69 and in 2002 between .57 and .71, with an average increment in grade level of only .18 for narrative and .59 for discursive writing.

Clearly, it is difficult to get exact agreement even using relatively short scoring rubrics and, thus, the consensus values are significantly more robust, using adjacent agreement. However, it is important that the scale is sufficiently long to support such an approach to obviate consensus through chance. Nevertheless, consensus estimates in the studies reviewed for exact agreement range typically between 40% and 60% and adjacent agreement rates of 80%–100%, are commonly found. Consistency coefficients of between .70 and .80 are reasonably consistently reported for standardized performance assessments in writing in which all students write responses to the same prompts. This is in contrast to portfolio writing samples that have consistency coefficients that often fall in the range .40 to .60. Measurement coefficients, reported much less often in the studies, tend to fall in the range .60 to .80 suggesting that a sizeable proportion of score variance is attributable to rater agreement. Thus, it appears that the accurate and consistent scoring of writing is a complex and difficult process but one which can be reasonably attempted provided the consequences of such assessment are relatively low-stakes.

3. Assessing writing in New Zealand

The Assessment Tools for Teaching and Learning (asTTle) Project is a New Zealand curriculum-based, voluntary use, teacher-controlled national assessment program released in 2003 as part of the Ministry of Education's assessment for learning strategy (Brown, 1998a; Brown & Hattie, 2003; Ministry of Education, 1994a, 1997). The asTTle is a software package distributed to schools for use as a classroom resource so that teachers can obtain feedback as to the learning needs of their own students. The asTTle software gives teachers the ability to customize tests according to the content and difficulty preferences they have for their own students. Teachers can create 40-minute, paper-and-pencil tests of reading, writing, or mathematics in either English or Maori languages and receive interactive, graphical reports that interpret student performance relative to both norms and criteria (Hattie, Brown, & Keegan, 2002).

The asTTle software is intended for use with students in Years 4–10 (aged 9–16) and all assessment items and tasks in asTTle Version 2 are indexed to Levels 2–4 of the eight progression levels described in the New Zealand English curriculum (Ministry of Education, 1994b). Note that asTTle Version 4 will extend this bank to Levels 2–6. The New Zealand English curriculum constitutes an inclusive, common framework to guide the instruction, assessment, and reporting of students' learning of written, oral, and visual language whether the students have English as first or additional language (Brown, 1998b) and thus all asTTle assessment tasks and rubrics are equally applicable within the New Zealand context for assessing English as L1 or L2 writing.

The curriculum levels are broad bands of progression each of which covers approximately 2 years of instruction. Because of the breadth of these progression

descriptors, the asTTle Project developed, with reference to current classroom practice, more closely described rating scales (Glasswell, Parr, & Aikman, 2001). The curriculum describes three types of writing (i.e., expressive, poetic, and transactional), which cover all the forms, purposes, and genres of writing. This breadth required further specification and so the asTTle Project developed a socio-communicative approach to writing that invoked social purpose as the basis for classifying kinds of writing. That approach also allowed identification of powerful dimensions of writing that lead to progress in the skill of communicating through written language (Glasswell et al., 2001). Thus, asTTle Project developed a series of six scoring rubrics (one each for persuade, instruct, narrate, describe, explain, and recount), each of which had criteria mapped to the broad characteristics identified in the New Zealand English Curriculum for Levels 2–4 (Ministry of Education, 1994b). Seven scoring variables or traits were identified for each writing rubric that related to either the deep (i.e., audience awareness and purpose, content inclusion, organisation or coherence, and language resources) or surface (grammar, spelling, and punctuation) features of writing. Additionally, multiple, age-appropriate writing prompts were developed with student-instructions intended to elicit the characteristics of writing associated with one of the socio-communicative purposes.

The asTTle scoring rubrics for writing were developed by a team of writing curriculum experts, and then refined through a workshop panel in which primary school teachers practiced scoring and commented on the clarity, appropriateness, and completeness of the scoring rubrics. The refined rubrics were then published (Glasswell et al., 2001) and used in the scoring of nationally representative samples of student writing (and specifically in the two studies reported in this paper). Each scoring trait has a rubric with three main grades which could be augmented by the raters resulting in a scale with 11 possible scores (i.e., below Level 2 Basic, 2 Basic, 2 Proficient, 2 Advanced, 3 Basic, 3 Proficient, 3 Advanced, 4 Basic, 4 Proficient, 4 Advanced, above 4 Advanced). The terms Basic, Proficient, and Advanced have been adopted as labels for the augmented scores to better represent progress through each level and refer to the early, middle, and late thirds of development within each level. The rubric for the deep features of persuasive writing as used in Study 2 is provided in Appendix A.

In order to generate national norms against which the performance of any teacher's students could be compared within the asTTle software, it was necessary to conduct national calibration exercises. Nationally representative samples of students in Years 5–8 (age 10–13) completed one assigned prompt each in a 40 minutes classroom period supervised by their own teacher. Students participated in a 10 minutes brainstorming and planning session before writing an individual single-draft piece of writing, although they were prompted by the teacher 5 minutes before the end of writing time to review their writing. These prompts were essentially impromptu activities which were scored as first drafts and thus may, as Davis, Scriven, and Thomas (1981) argue, only represent students' ability to create, and organise their thoughts under some time pressure albeit with a modicum

of assistance. It should be noted that when teachers use the asTTle software they are able to select prompts and purposes to suit their own teaching objectives and are encouraged to provide adequate preparation prior to the administration of the assessment proper.

Two studies have been conducted into the reliability of asTTle writing scores. The first study was a centrally run marking panel in which nationally representative samples of student writing in Years 5–8 were scored. These data were used in the asTTle software as part of the representative norms against which teachers using the asTTle software could compare their own students' achievement. The second study was a small intervention study intended to discover whether low achieving students could be taught to improve the quality of their persuasive writing. As demonstration of effectiveness of the intervention, a reliability study was conducted. The paper will comment on the aspects of the asTTle writing assessment program that contributed to the accuracy results obtained.

4. Study 1

In accordance with most large-scale assessments of writing (Dyson & Freedman, 1991), a marking panel to score six different asTTle writing prompts involving two different writing purposes (i.e., instruct and describe) was conducted in Auckland for a week in January, 2002. There was an average of 589 (SD = 269) different scripts to be marked for each of the six writing prompts. Marking such a range and number of scripts in a relatively short period has additional challenges for training and maintaining marker confidence and reliability. Seventeen experienced classroom teachers were recruited by the University of Auckland on the recommendation of the second author, a New Zealand ex-primary school teacher. None of the markers were experienced in large scale marking operations though some had been involved in previous asTTle development and review workshops. Because instructional feedback to teachers and students is a major goal of the asTTle Project, an analytic approach to scoring each of the seven traits was taken. This approach required each marker to read and score the script for each trait separately to reduce scores for each trait influencing the rating of another (Cooper, 1984).

Several procedures were used to ensure and monitor the quality, accuracy, and consistency of scoring. A half-day training program, checking of the number of scripts marked per hour, cross-checking or moderation of scoring by expert markers, and the use of control or reliability-checking scripts were used as measures of quality. The training of markers involved a 1.5-hour grammar instruction lecture on the first day, an overviewing of the scoring rubrics, and specific training on the progress indicators for describe/report writing for an hour on the first day. The second day provided a review of the describe/report writing progress indicator for a further hour. Before each writing task was started, detailed and specific training was provided clarifying the task and its rubric for approximately 15–20 minutes.

Before each writing task was marked, sample scripts were introduced, discussed, and used as reference benchmarks for subsequent marking. Subsequent training on the second writing purpose rubric (i.e., instruct) for about an hour was undertaken once all the tasks related to describe/report purpose writing were completed. In total, training time for this marking panel was around 4 hours out of 25 hours paid employment.

The rate of scoring completion was monitored throughout the panel. Allowing for initial and retraining periods, the marker rate averaged 7.2 scripts per hour. Some variation in speed occurred but overall the marking rate for the use of this instrument was good. As with most marking operations, there was some variability in marker rate. The rate of script marking achieved may have been slower than previous marking panels but two factors affected completion rate; specifically, changing task five different times and the relative lack of experience and skill the markers had in large-scale marking. No panellists were discontinued for overly slow scoring.

Cross-checking or moderation by a chief examiner or team leader (Alderson, Clapham, & Wall, 1995) was used to evaluate consensus of scoring between markers. Two scripts from each bundle of 20 marked by each marker were cross-marked by one of two expert markers. Margaret Aikman, a co-developer of the asTTle writing progress indicators and an experienced primary school teacher trainer, assisted the second author in the cross checking of the scores. Feedback on marking was provided to all markers following the cross-marking. Additional guidance was provided to enhance the accurate marking of markers (between 0 and 3 out of 17 each day) identified as being less consistent overall or in some score areas (between 2 and 3 of the 7-score variables each day). No markers were discontinued for inaccuracy, though several had additional numbers of scripts cross-checked by the panel's two expert markers to ensure greater reliability in non-control marking. The areas that the teachers needed most extra instruction in were grammar and language resources (e.g., sentence structure, complex sentences, and punctuation), marking against criteria, marking within curriculum levels using the sub-levels of Basic, Proficient, and Advanced, and understanding the required content for the prompts.

Scoring of a common 'reliability script' (Alderson et al., 1995) and statistical analysis of scoring to further evaluate consistency and consensus of scoring was conducted four times throughout the workshop. One control script was issued per marking day (beginning on the second day of the marking panel after initial training). These scripts were selected from unmarked scripts for the day and were issued without amendments or annotations. Control script information was used to provide information about the nature of specific redirection required by individuals or groups with feedback on each control script being given immediately after lunch each day. Thus, the data available for analysis consisted of four scripts marked by each of the 17 panel markers and with seven different scores for each script (one each for the seven traits of audience, content, organisation, language resources, grammar, punctuation, and spelling).

4.1. Analysis

Inter-rater reliability was calculated, using consensus, consistency, and measurement approaches. Specifically, the following statistics were obtained: (a) the average percentage of adjacent agreement for the seven score variables for each day's common script between the 17 panel markers and the scores of the second author, (b) the coefficient alpha for the seven score variables assigned by the 17 scorers to each day's common script, and (c) the dependability phi correlation for the seven score variables assigned by the 17 scorers to each day's common script. In the first case, agreement was calculated as having occurred if markers' scores were either the same as or within one score to the marks awarded by the expert marker. This meant correct scores had to fall within the range of +1 or −1 of the assessment leader's score on the 11-point scale. Close agreement rates of the 17 teachers to the scores assigned by the expert marker for the seven scores per script ranged from 66% to 92% with an average of 75% (Table 1).

For consistency, the seven scores for each task from the common script for each day were analyzed using the SPSS coefficient alpha reliability routine (Table 1). The alpha coefficient estimates of reliability for each day ranged from .53 to .87. Note that the lowest average coefficient came about through the classic effect of high agreement causing reliability deflation (i.e., one score variable for the third test had 100% consensus agreement among all raters). For measurement approach estimates, the Brennan and Kane dependability index (φ) was calculated by obtaining the between-raters effect error mean square and dividing it by the sum of the absolute error variance of the set of ratings and itself using the formula $\varphi = \sigma_p^2/(\sigma_p^2 + \sigma_{ABS}^2)$ (Shavelson & Webb, 1991). Values ranged from .67 to .95 with an average of .77, indicating that for the most part the observed scores approach closely the threshold for dependable values (Table 1).

Taken together, the three approaches show, based on 17 raters scoring one daily common script for seven writing traits (i.e., four deep characteristics and three surface characteristics) on an 11-point scale, that there was substantial agreement

Table 1
Control script average inter-rater reliability calculations by day

Day	Approach (statistic)		
	Consensus (percent adjacent agreement)	Consistency (alpha coefficient)	Measurement (phi dependability coefficient)
1	.66	.87	.67
2	.72	.83	.95
3	.92	.53	.76
4	.71	.77	.69
Average	.75	.75	.77

Note: Each statistic is based on 17 raters, giving one common script scores for seven dimensions of writing.

about: (a) the mean score or curriculum level for each script, (b) the classification of scripts as high or low, and (c) the source of the scores. The results for the first common script show acceptable levels of consistency agreement after the panellists had received only one half-day training session. After a second day of marking the consensus and measurement approaches also show robust levels of inter-rater reliability. These figures indicate that the accuracy of the scoring given a half-day of training is such that they can be used as a reliable set of norms for the classroom teachers who use the asTTle software. Study 2 examines, in part, the impact of light- or zero-training conditions on the accuracy of scoring using the asTTle rubrics.

In addition to establishing the consistency of the marking, evaluative comments collected from markers indicated that the workshop provided many benefits. All markers agreed that their work was well explained and directed, and that the marking instructions were clear. They all agreed that the training had given them a clearer understanding about curriculum levels in writing and felt that the rubrics used to mark the scripts were consistent with curriculum Levels 2–4 of the English curriculum and bands within those levels. All markers agreed that participation in the workshop would help them better plan for teaching writing and for assessing students' writing better.

5. Study 2

The asTTle rating scales and prompts were designed to meet the progress characteristics of curriculum Levels 2–4. Although, norms were generated for students in Years 5–7, Harland (2002) found that students who were predominantly from language backgrounds other than English and of minority group ethnicities, in low socio-economic secondary schools exhibited low levels of writing skill upon entry to Year 9. Harland devised an intervention study using the asTTle scoring rubrics that taught low SES high school students to write persuasively. Just over 60 students, in four classes in three schools, completed an argumentative writing training program involving explicit instruction and modelling, as well as training in meta-cognitive monitoring and scaffolding (i.e., gradually removed supportive instruction) of writing.

Across the four classes an average of 27.25 hours ($SD = 3.3$) was spent teaching the students how to write persuasively; in other words, given a 4-hour per week schedule typical of high schools, the program lasted an average of 7 weeks per class. In addition, 2 hours on average were spent conducting assessments across the program. As part of the training program students were asked to complete a series of four different persuasive essay topics. All opinion-based persuasive or argumentative essays were scored by the third author according to the asTTle persuasive purpose progress indicators for four deep score traits exhibited in Appendix A and the three surface scores (i.e., grammar, punctuation, and spelling). Additionally, an approximate word count of the pre-test and the second during-instruction essay were taken.

A consistency estimate of reliability for the essays in Harland's study (63 students by four essays) was very high: $\alpha = .95$ for total score, $\alpha = .92$ for deep score, and $\alpha = .93$ for surface score. These estimates indicate the researcher exhibited very consistent intra-rater scoring patterns. More interestingly, a sample of 42 scripts were cross-checked by a panel of three writing specialists who were given no face to face instruction. These external markers were given just the asTTle progress indicators shown in Appendix A and a one-page guide to for scoring. Consistency coefficients of agreement between the researcher and the three expert markers averaged $r = .70$ ($SD = .05$) for deep scores and $r = .60$ ($SD = .07$) for surface scores. These consistency estimates provide an indication of the floor level of inter-rater reliability in a zero-training condition.

The more important finding of this study was the large gain in the quality of persuasive writing in such a short training program. Initially, the bulk of students scored, according to the asTTle progress indicator for deep features, between Level 2 Proficient and Level 3 Basic (M = Level 2 Advanced; $SD = 1.5$ curriculum sub-levels). The average length of their first essay was estimated at 180 words ($SD = 59$). The essay submitted mid-way through the program (at the end of the second or third instructional unit depending on class) showed significantly improved average results for deep scores. The deep scores had an effect size of 1.6, reflecting an average improvement of one whole curriculum level. The score range for one standard deviation either side of the mean was Level 3 Basic to Level 4 Basic (M = nearly Level 3 Advanced; $SD = 1.4$ curriculum sub-levels). The average essay length increased to 221 words ($SD = 7.3$) representing an increase in elaboration and reasoning. Some 2 months after the completion of the intervention, another argumentative essay was supplied by the students. The score range for one standard deviation either side of the mean was Level 3 Basic to Level 3 Advanced (M = Level 3 Proficient; $SD = 1.1$ curriculum sub-levels). This clearly indicates that the gains achieved through this program were largely sustained.

Another way to evaluate the impact of this intervention is to compare it to the cross-sectional national year norms for writing found in the asTTle writing standardization (Hattie et al., 2002). The average score for persuasive writing in Years 4–6 falls within Level 2 Advanced, Year 7 within 3 Basic, and Year 8 in Level 3 Proficient. Thus, the observed mean gain from 2 Advanced to 3 Proficient for the Year 9 (i.e., first year secondary school) students in Study 2 constituted a gain equivalent to more than three years' average progress.

6. Discussion

Both of these scoring studies demonstrated that acceptable reliability in scoring of extended writing tasks for use in the low-stakes environments of classrooms could be obtained with New Zealand primary teachers who had had little previous experience in mass marking. The consensus rates, consistency coefficients, and

measurement coefficients reported here are similar to those found in other jurisdictions where standardized large-scale assessments are commonplace. A reasonably small amount of training and monitoring was needed to ensure consistency across a panel of raters; and the .70 agreement between the third author and the triplet of untrained raters suggests that the asTTle rubrics provide relatively robust guidance for the low-stakes uses for which they were designed. The results, nevertheless, suggest that classroom teachers may benefit from a 2-day training package consisting of orientation to the socio-communicative approach and practice marking so that classroom use of asTTle scoring rubrics and prompts approaches the degree of accuracy found in a centrally scored national assessment. Furthermore, these studies show that the teachers can have confidence in the norms underlying the asTTle tools because of the accuracy of the scoring reported in Study 1. These studies also provide guidance to schools and teachers as to how they can conduct not only similar high quality school-based assessment of writing, but also create writing intervention programs.

Acknowledgements

The authors acknowledge the funding of the New Zealand Ministry of Education to the Assessment Tools for Teaching and Learning Project. The research in this report was initially reported in two asTTle technical reports which are available from http://www.asttle.org.nz.

References

Airasian, P. W. (1997). *Classroom assessment* (3rd ed.). New York: McGraw-Hill.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council for Measurement in Education (NCME). (1999). *Standards for Educational & Psychological Testing*. Washington, DC: American Educational Research Association.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (ETS Research Report No. 99-3). Princeton, NJ: Educational Testing Service.

Brown, G. T., & Hattie, J. A. (2003). *A national teacher-managed, curriculum-based assessment system: Assessment tools for teaching & learning (asTTle)* (asTTle Tech. Rep. No. 41). Auckland, NZ: University of Auckland/Ministry of Education.

Brown, G. T. L. (1998). Assessment in English. *English in Aotearoa*, *36*, 62–67.

Brown, G. T. L. (1998). The New Zealand English curriculum. *English in Aotearoa*, *35*, 64–70.

Coffman, W. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English*, *5*, 24–36.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

College Board. (2001). *AP technical corner: Scoring the free-response section*. Retrieved September 15, 2001. From http://www.collegeboard.com/ap/techman/chap3/scorefc.htm.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research* (ETS Research Report No. 84-12). Princeton, NJ: Educational Testing Service.

Davis, B. G., Scriven, M., & Thomas, S. (1981). *The evaluation of composition instruction*. Point Reyes, CA: Edgepress.

Dyson, A. H., & Freedman, S. W. (1991). Writing. In: J. Flood, J. M. Jensen, D. Lapp, & J. R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 754–774). New York: Macmillan.

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. K. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Tech. Rep. No. 337). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Gearhart, M., Herman, J. L., Novak, J. R., Wolf, S. A., & Abedi, J. (1994). *Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric* (CSE Tech. Rep. No. 389). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Glasswell, K., Parr, J., & Aikman, M. (2001). *Development of the asTTle writing assessment rubrics for scoring extended writing tasks* (asTTle Tech. Rep. No. 6). Auckland, NZ: University of Auckland/Ministry of Education.

Green, S., Johnson, M., O'Donovan, N., & Sutton, P. (2003, July). *Changes in key stage two writing from 1995 to 2002*. Paper presented at the United Kingdom Reading Association Conference, Cambridge, UK.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Harland, D. (2002). *Teaching argumentative writing to Year 9 students*. Unpublished Master of Arts thesis. University of Auckland, Auckland, NZ.

Hattie, J. A., Brown, G. T. L., & Keegan, P. J. (2002). *Assessment Tools for Teaching and Learning (asTTle) manual: Version 2, 2003*. Wellington, NZ: Learning Media.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1* (3), 201–224.

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

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1993). *Interim report: The reliability of Vermont portfolio scores in the 1992–1993 School Year* (CSE Tech. Rep. No. 370). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CSE Tech. Rep. No. 355). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Lee, Y. W. (2001). *The essay scoring and scorer reliability in TOEFL CBT*. Paper presented at the Annual Meeting of National Council on Measurement in Education (NCME), Seattle, WA.

LeMahieu, P., Gitomer, D., & Eresh, J. (1995). Large-scale portfolio assessment: Difficult but not impossible. *Journal of Educational Measurement: Issues and Practice, 14*, 11–28.

Linn, R. L. (1991). *Cross state comparability of judgments of student writing: Results from the New Standards Project* (CSE Tech. Rep. No. 335). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and evaluation in teaching* (8th ed.). New York: Macmillan.

Linn, R. L., Kiplinger, V. L., Chapman, C. W., & LeMahieu, P. G. (1992). Cross-state comparability of judgments of student writing: Results from the New Standards Project. *Applied Measurement in Education, 5* (2), 89–110.

Longford, N. (1994). *A case for adjusting subjectively rated scores in the advanced placement tests* (Tech. Rep. No. 94-5). Princeton, NJ: Education Testing Service.
McMillan, J. H. (2001). *Classroom assessment: Principles and practice for effective instruction* (2nd ed.). Boston, MA: Allyn & Bacon.
Ministry of Education. (1994). *Assessment: Policy to practice*. Wellington, NZ: Learning Media.
Ministry of Education. (1994). *English in the New Zealand curriculum*. Wellington, NZ: Learning Media.
Ministry of Education. (1997). Assessment for better learning. Curriculum development: A report on national assessment developments. *Update*, *18*.
Novak, J. R., Herman, J. L., & Gearhart, M. (1996). *Issues in portfolio assessment: The scorability of narrative collections* (CSE Tech. Rep. No. 410). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing*, *7*, 143–164.
Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6th ed.). Boston: Allyn & Bacon.
Resnick, L., Resnick, D., & De Stefano, L. (1993). *Cross-scorer and cross-method comparability and distribution of judgments of student math, reading, and writing performance results from the New Standards Project Big Sky Scoring Conference* (CSE Tech. Rep. No. 368). Los Angeles, CA: UCLA Graduate School of Education, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment Research & Evaluation*, *9* (4).
Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K., & Powers, T. (1999). A generalisability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology*, *20* (2), 107–127.
Supovitz, J. A., MacGowan, A., III, & Slattery, J. (1997). Assessing agreement: Interrater reliability of portfolio assessment in Rochester, New York. *Educational Assessment*, *4* (3), 237–259.

Appendix A. asTTle persuade progress indicators by curriculum level

Dimension	Level 2	Level 3	Level 4
Audience awareness	Evidence that writer recognizes that his/her opinion is needed. May use language to state opinions from a personal perspective.	Language use and writing style generally appropriate to audience. Writer states his/her position on the issue and makes some attempt to influence.	Language use and writing style appropriate and directed to audience (e.g., writing attempts to persuade reader). Clearly stated position is evident and maintained throughout.

Appendix A (*Continued*)

Dimension	Level 2	Level 3	Level 4
Content/ideas	Writing covers some (1 or more) task and topic appropriate domains: (e.g., position statement—writer identifies position on the issue, makes 2 or more simple opinion/statements related to the topic, makes use of a final statement to round off the text in some way). Can include many statements unrelated to the topic and/or task.	Most argument domain elements are present (main points, some supporting evidence/illustration, and re-statement of position). Some elaboration of main points occurs. May include information that does not contribute to argument.	Argument domain elements (e.g., position statement, main points, illustration/evidence of main points, re-statement) are comprehensive and elaborated. Content is relevant and functions to add weight to the writer's position.
Structure/ organisation	Semblance of organisation is evident (e.g., some grouping of ideas). But text may be limited because of presentation of opinion statements as discrete elements.	Evidence of attempts at overall structuring of content through grouping ideas within and across sentences (may use devices, such as paragraphing and simple linking of ideas through conjunctions, such as because, and, since, although, etc).	Content managed effectively through grouping and/or paragraphing main ideas and supporting evidence. Ideas are linked in more complex ways (e.g., varied use of linking words and phrases, conjunctions and text connectives e.g., on the one hand, however, although).

Appendix A (*Continued*)

Dimension	Level 2	Level 3	Level 4
Language resources	Language has structure of simple opinion statements (e.g., may be stated from a personal perspective "I reckon"). Topic related language present but little opinion is conveyed through language choices (e.g., nouns may be neutral, may have basic descriptors, or may lack simple adjectivals. Verbs and adverbials may be limited). Shows some understanding of the use of pronouns but reference (the who or what) may be unclear or overused. Simple sentences used but may attempt complex sentences.	Evidence of use of some task appropriate structures and language (e.g., attempts to use verbs in passive structures to make arguments seem more objective and convincing). Evidence that the writer is a beginning to select language to create a particular effect and to influence the reader (e.g., "point of view" nouns, viewpoint adverbials, opinion adjectives, adverbs and adjectives to add detail and weight to opinion statements and evidence, some use of modal auxiliary verbs (can, might, should, may) present). May be some unclear or repetitious reference. Many simple sentences correct. Some successful complex sentences evident.	Consistent use of appropriate language for task and topic (passive structures may be used to make the argument seem more formal and objective, modal auxiliaries (may, might, can, should, shall) may be used to add persuasive power). Language supports a particular viewpoint and is used to persuade the reader (e.g., "point of view" nouns, viewpoint adverbials, opinion adjectives, adverbs, and adjectives to add detail and weight to opinion statements and supporting evidence). Reference links clear (pronoun use). Most sentences correct. Control of complex sentences evident, where appropriate. Uses complete sentences.