Technical Report # 26
Assessment Tools for Teaching and Learning

ACCURACY IN THE SCORING OF WRITING:

STUDY IN LARGE-SCALE SCORING OF asTTle WRITING ASSESSMENTS

Submitted by the Assessment Tools for Teaching and Learning team,

Auckland UniServices Ltd

University of Auckland

August 2003

ACCURACY IN THE SCORING OF WRITING:

STUDY IN LARGE-SCALE SCORING OF asTTle WRITING ASSESSMENTS

asTTle is funded by the Ministry of Education to Auckland Uniservices at the University of Auckland to research and develop an assessment application for Reading, Writing, Mathematics, Pānui, Pāngarau, and Tuhituhi for Years 4-12 (Levels 2-6) for New Zealand schools. We acknowledge this funding, and thank the Ministry of Education for their continued assistance in the development of this project.

This report reviews the work of a four-day asTTle writing marking panel in which six writing tasks were scored by 17 markers. The processes used to ensure reliable scoring, and the degree of accuracy obtained through those procedures are reported. An average 75% exact score marking and an average dependability of .77 were found. These levels of agreement provide sufficient basis for having confidence in the norms underlying the asTTle writing assessments for use in low-stakes classroom assessment. The processes also provide guidance to schools as to how they can conduct high quality school-based assessment of writing.

Dr Kathryn Glasswell, then of University of New South Wales, was instrumental in developing the refined scoring rules, led the 2002 marking panel, and provided a report on that panel. Dr Judy Parr, University of Auckland, and Mrs Margaret Aikman, Auckland College of Education assisted her work on refined scoring rules. Dr Gavin Brown, University of Auckland, analysed the reliability study data and wrote this report. The project team would like to acknowledge the assistance of Mr Jeremy Zwiegelaar and Miss Angela Parker who helped run the various scoring rubric development and scoring panel workshops.

John Hattie
Project Director, asTTle
August, 2003

The bibliographic citation for this report is:

Glasswell, K., & Brown, G. T. L. (2003, August). *Accuracy in the scoring of writing: Study in large-scale scoring of asTTle writing assessments.* asTTle Technical Report 26. University of Auckland/Ministry of Education.

The accurate scoring of writing is a necessary 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 (Linn & Gronlund, 2000; Popham, 2000); cross-checking or moderation of marking (Gronlund & Linn, 1990); systematic scoring processes (Airasian, 1997; McMillan, 2001); training of markers (AERA/APA/NCME, 1999), and the use of panels (AERA/APA/NCME, 1999). Inter-rater correlation is one of the accepted techniques for establishing the consistency or reliability of scoring with values exceeding .70 indicating acceptable similarity of scoring for low-stakes classroom uses. Correlations of .80 or higher support high-stakes interpretations based on the obtained scores for writing.

Project asTTle developed a socio-communicative approach to writing that invoked social purpose as the basis for classifying kinds of writing and identifying powerful dimensions of writing that lead to progress in the skill of communicating through written language (Glasswell, Parr, & Aikman, 2001). That approach developed a series of six scoring rubrics (one each for persuade, instruct, narrate, describe, explain, and recount), each of which had criteria mapped to Levels 2—4 of the New Zealand English Curriculum (Ministry of Education, 1994). Seven scoring variables were identified for each writing rubric (i.e., audience awareness & purpose, content inclusion, organisation or coherence, language resources, grammar, spelling, and punctuation). The scoring rubrics were developed initially by a team of writing curriculum experts and were 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 and used in the scoring of nationally representative samples of student writing. Each scoring variable has a range of 11 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).

This report describes the processes used in one marking panel to ensure reliable scoring and the degree of accuracy obtained through those procedures. It provides sufficient basis for having confidence in the norms underlying the asTTle writing assessments and provides guidance to schools as to how they can conduct high quality school-based assessment of writing.

Procedures For Large-Scale Scoring Of Writing

A marking panel to score six different asTTle writing tasks (numbered 25-30) involving two different writing purposes (i.e., instruct and report or describe) was conducted in Auckland for a week in January 2002. Marking such a variety of tasks 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 Kath Glasswell, Judy Parr, and Margaret Aikmann. None of the participants were experienced in large scale marking operations though some had been involved in previous asTTle development and review workshops.

Several procedures were used to ensure and monitor the quality, accuracy, and consistency of scoring. Some training, checking of the number of scripts marked per hour, crosschecking 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 on the first day involved a lecture on understanding grammar for scoring, over viewing of the scoring rubrics, and specific training on the progress indicators for report or describe writing. The second day provided a review of the report or describe 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 or report purpose were completed. In total, training time for this marking panel was around four hours out of 25 hours.

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 achieved may have been slower than other 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.

Crosschecking or moderation by expert assessors was used to ensure consistency of scoring between markers. Two scripts from each bundle of twenty 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 first author in the cross checking of the scores. Feedback was provided to all markers re cross marking. Additional guidance was provided to enhance the accurate marking of markers (between zero and three out of 17 each day) identified as being less consistent overall or in some score areas (between two and three of the seven score variables each day). No markers were discontinued for inaccuracy, though several had additional numbers of scripts cross checked by Kath Glasswell and Margaret Aikmann to ensure greater reliability in non-control marking. The areas that the teachers needed the 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 required content for a task.

Establishing The Reliability Of Scoring

Scoring of common scripts and calculation of agreement correlation was conducted four times to establish reliability of scoring. 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.

Inter-rater reliability was calculated in three ways; (a) the percentage of agreement with the first author, (b) coefficient alpha between scorers across scripts, and (c) a dependability study. In the first case, agreement was calculated as having occurred if markers' scores were either the same as or adjacent 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).

Table 1.
Control Script Reliability Calculations

Day	Average Percent Agreement (*N*=17)
1	.66
2	.72
3	.92
4	.71
Average	.75

For the seven assessment items in each day's crosschecking script, the dependability of rater scoring was measured using the Brennan and Kane Dependability Index (Table 2). The Brennan and Kane Dependability Index (φ) is calculated by obtaining the *between-subjects effects error mean square* and dividing it by the sum of the *absolute error variance of the set of ratings* and itself: $\varphi = \sigma^2_p / (\sigma^2_p + \sigma^2_{ABS})$ (Shavelson & Webb, 1991). Values close to or greater than .80 are considered dependable and are seen in the marking of common scripts on the second and third day. However, the average φ = .77 across the four days is close to the critical threshold for dependable rating.

Table 2
Calculation of the Brennan and Kane Dependability Index (φ) Results

Day	σ^2_i	σ^2_p	$\sigma^2_{pi,e}$	Σ^2_{ABS}	φ
Tuesday	33.01	1.13	9.96	4.88	0.67
Wednesday	1.11	1.04	6.04	0.31	0.95
Thursday	0.63	0.19	0.36	0.12	0.76
Friday	17.57	1.36	6.06	2.70	0.69

It is worth noting that these reliability and dependability indices indicate the positive impact of the training and monitoring. Harland (2002), using the asTTle persuade scoring rubric, found inter-rater agreement between himself and three other markers who had received no training averaged .70 on deep scores. Thus, the scoring of writing for the creation of asTTle's background norms is sufficiently consistent with variance being attributed largely to agreement of scorers rather than error.

In addition to establishing the consistency of the marking, evaluative comments collected from teacher participants 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 New Zealand Curriculum Levels 2-4 (Ministry of Education, 1994) and bands within those levels. All markers agreed that participation in the workshop

would help them better plan for teaching writing and improve their assessment of student's writing. Markers agreed that the exercise had been an intensive professional development experience and would provide them with totally new ways of assessing writing in their classrooms.

Conclusion

This marking workshop demonstrated that good marking rates and 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. Evidence presented here and in Harland (2002) suggests that when teachers use asTTle scoring guidelines in classroom assessment the degree of accuracy will fall between .70 (no training) and .80 (half-day training). This is a level of accuracy and consistency suitable for establishing valid classroom instructional decisions.

References

AERA/APA/NCME. (1999). *Standards for Educational & Psychological Testing*. American Educational Research Association. Washington, D.C.

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

Glasswell, K., Parr, J., & Aikman, M. (2001). *Development of the asTTle Writing Assessment Rubrics for Scoring Extended Writing Tasks.* asTTle Technical Report 6. University of Auckland/Ministry of Education.

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 Masters thesis, University of Auckland, NZ.

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

McMillan, J. H. (2001). *Classroom assessment: Principles and practice for effective instruction.* (2nd. ed.). Boston, MA: Allyn & Bacon.

Ministry of Education. (1994). *English in the New Zealand curriculum.* Wellington, NZ: Learning Media.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. (6th ed.). Boston: Allyn & Bacon.

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