Do rubrics help to inform and direct teaching practice?

Stephen Humphry

University of Western Australia

Stephen Humphry is an Associate Professor with the Graduate School of Education at the University of Western Australia. He teaches masters units in Educational Assessment, Measurement and Evaluation and is involved in a number of research projects. He currently holds an Australian Research Council grant entitled *Maintaining a Precise Invariant Unit in State, National and International Assessment* with Prof David Andrich of UWA. He is a member of the Curriculum Council's Expert Measurement and Assessment Advisory Group and is involved in research on assessment and measurement more broadly in Western Australia and Australia. He has presented at international conferences and has visited and worked with international organisations and institutions, including MetaMetrics and the Oxford University Centre for Educational Assessment.

Dr Humphry completed his PhD under Professor David Andrich, with a focus on maintaining a common unit in measurement in the social sciences. His doctoral research involved advancements in item response theory as well as applied work to demonstrate the advancements lead to improved test equating. Prior to 2006, he worked for a number of years in industry as the Senior Psychometrician for the Department of Education Western Australia. During that time, he was responsible for the measurement and statistical analysis of data obtained in large-scale State testing programs. He designed and coordinated research and development projects associated with the assessment program, as well as projects focusing on the use of student data for monitoring and evaluating student performance.

Dr Humphry has several lines of active research, the most central being work on developing a general framework for defining and realizing units in the social sciences. His work in education has included research on: test equating; rubrics; applications of the process of pairwise comparison; and teacher effectiveness. He is also pursuing research on parallels between biological and cognitive growth that mirror parallels between methods of data analysis used by Sir Julian Huxley and Georg Rasch.

Sandra Heldsinger

University of Western Australia

Sandy Heldsinger has worked at the University of Cambridge Local Examination Syndicate (UCLES) as a research officer responsible for establishing programs of trialing and pre-testing, as project coordinator for the Australian National Benchmarking Equating Study and as an associate lecturer at Murdoch University in educational assessment. She worked as Senior Educational Measurement Officer, Population Testing in Department of Education, WA for over seven years and her work included coordination of random sample assessment programs of student achievement in the social outcomes of schooling and the society and environment learning area; and the coordination of the annual, full cohort WA assessment program.

In her work with the Western Australia Department of Education, Dr Heldsinger conceptualised and led the development of a suite of publications that assist teachers to interpret the data from system level assessment programs and to understand the frameworks that guide teaching and assessment. Dr Heldsinger commenced as a Lecturer, UWA in 2006 where she teaches in assessment and educational measurement.

Background

Assessment in learning domains that require an extended performance of some kind (for example, an essay or work of art) has been considerably more vexed than for domains where closed response items, such as multiple-choice items or short answer items, are valid. Different countries have grappled with the issues related to performance assessment in slightly different ways depending on the dominant assessment regime, but the underlying issues remain very similar. In the United Kingdom (UK), for example, the assessment of a single composition in a fixed-time examination, marked by a detailed marking scheme, is seen as the archetypal assessment that has influenced practice in the current assessment regime (Wilkinson et al., 1980). In the 1930s, dissatisfaction with this way of marking led to a debate about analytical marking as opposed to impressionistic marking, where analytic marking consisted of a series of headings or criteria and an allocation of marks available for each criterion (Wilkinson et al., 1980). Concerns that this way of marking did not result in the best essay obtaining the top mark led to an exploration of impression marking, where the markers were provided with a small number of criteria to consider when marking; but rather than being provided with a mark for each criterion, they arrived at a judgment of an overall mark.

In the 1980s there was a renewed interest in performance assessment. In part, this renewed interest resulted from the imposition in some countries, principally the United States of America (USA), of system-level standardised assessments where the predominant question format was multiple choice or short answer. Performance assessments were considered to be an integral aspect of educational reform because of their capability of measuring learning that could not be assessed through the more closed response formats, and because of their value for curricular and instructional changes (Lane & Stone, 2006).

It appears that the renewed interest in performance assessment coincided with educational reform that was happening in a number of countries. This reform saw a move away from syllabus documents which provided details of what teachers needed to teach, to frameworks that described progression in student learning. In the UK, this framework took the form of the *National Curriculum;* in Australia, *National Profiles* were developed and these in turn were reworked by each State educational authority. In Western Australia, the framework was referred to as the *Outcomes and Standards Framework*. In 1995, Spady (cited in Dimmock, 2000) outlined the features of Outcome-Based Education, two of which were:

- Schools define and communicate to students and parents the performance criteria and standards that represent the intended learning and outcomes expected
- Assessment is matched to the criteria and every student is eligible for high marks.

Outcome-based education has the same intentions as rubrics: to capture the essence of student performance or development at various levels.

When the difficulties experienced in assessing performances is considered in relation to the move towards defining performance criteria and standards it is not surprising that rubrics have become so popular. But are they as Popham (1997) suggests 'instructionally fraudulent'? Do rubrics help to inform and direct teaching practice?

To explore these questions further, this presentation firstly considers the typical rubric structure. It then provides an overview of a series of extensive empirical studies of the assessment of students' narrative writing. This presentation focuses on the qualitative research. The quantitative research undertaken is reported separately (Humphry & Heldsinger, 2009). Finally the implications of the findings from these studies for use of rubrics as instructional tools are discussed.

Overview of rubrics

A scoring rubric typically has three parts: (1) performance criteria (2) performance level and (3) a description of features evident in the performance level. The performance criteria are related to the task; so for example if a teacher was assessing his or her students' skills in devising an advertising brochure, one of the criterion could be the *visual appeal* of the brochure. The performance levels may be indicated by the labels *weak, good, very good and outstanding* or by using numbers to indicate increasing levels of achievement. The descriptions that accompany each of the performance levels summarise in some way the features of the performance at that level.

The predominant format of rubrics is that each criterion has the same number of performance levels, and most commercially available rubrics have four performance levels for each criterion. We will now focus on a specific example to examine these features of rubrics and the implications for using rubrics to inform and direct teaching practice.

Rubric for the assessment of narrative writing

The rubric discussed here was devised to assess narrative writing in the full-cohort testing program in Western Australia. The rubric was extracted from the Western Australian *Outcomes and Standards Framework* (OSF). The OSF describes the typical progress students make in each of eight learning areas. Learning in these areas is described in terms of eight stages, referred to as eight levels. This rubric consisted of nine criteria. Markers were required to make an on-balance judgment as to the level (1–8) of each student's performance overall and then they were required to assess each performance in terms of *spelling, vocabulary, punctuation, sentence control, narrative form of writing, text organisation, subject matter, and purpose and audience.*

The category descriptions within each criterion were derived directly from the OSF. That is, the description used to determine a score of 2 in spelling was taken directly from the description of the level 2 performance in the OSF; the description for a score of 3 was taken directly from the level 3 description in the OSF, and so on. The number of categories for each criterion is shown in Table 1.

Several interrelated issues with the psychometric properties of the data obtained from this assessment were identified, the most tangible being the distribution of student raw scores.

Figure 1 shows the raw score distribution of Years 3, 5 and 7 students in 2001, 2003 and 2004. It can be seen, firstly, that the distributions remained relatively stable over the period (2001–2004). This stability was achieved through the training of markers and in particular through the use of exemplar scripts, rather than by applying post-hoc statistical procedures.

Secondly, and most importantly, the graph shows that although there is a large range of possible score points (1– 61), the distribution clusters on a relatively small subset of these (in particular, around scores 18, 27 and 36).

Table 1: Original classification scheme for the assessment of writing

Aspect	Score Range	Aspect	Score Range
On-balance judgment (OBJ)	0 – 8	Form of Writing (F)	0 – 7
Spelling (Sp)	0 – 5	Subject Matter (SM)	0 – 7
Vocabulary (V)	0 – 7	Text Organisation (TO)	0 – 7
Sentence Control (SC)	0 – 7	Purpose and Audience (PA)	0 – 7
Punctuation (P)	0 – 6		
		Total score range	0 – 61

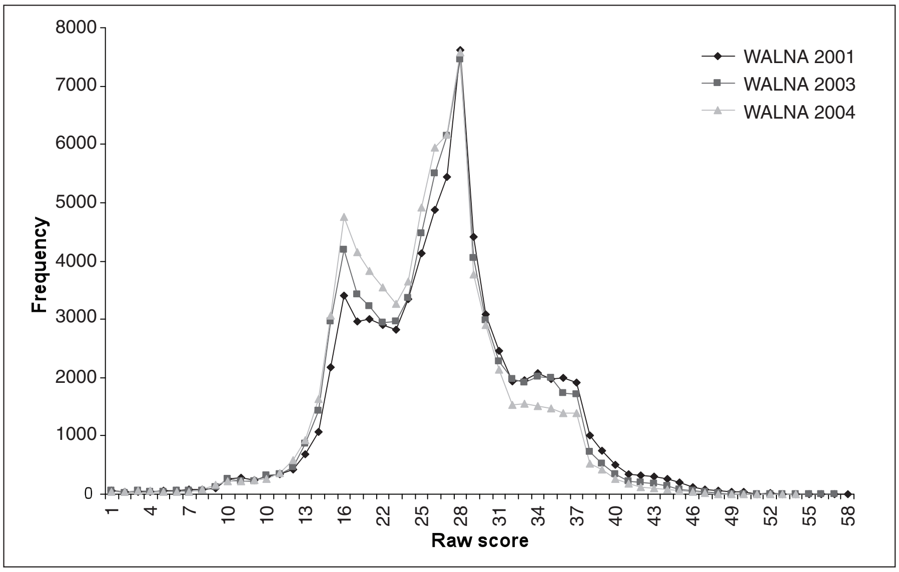

Figure 1: The raw score distribution of Years 3, 5 and 7 students' narrative writing as assessed through the Western Australian Literacy and Numeracy Assessment in 2001, 2003 and 2004

Examination of logical and semantic overlap in the rubric

A close analysis of the rubric revealed logical and semantic overlap in some of the performance criteria and levels. Table 2 shows an extract taken from the rubric and it can be seen that a student who writes a story with a beginning and a complication would be scored 2 for the criterion, *form of writing*. This student will necessarily have demonstrated some internal consistency of ideas (category 2, *subject matter*). Similarly if a student has provided a beginning and a complication, he or she has most probably provided a narrative that contains two or more related connected ideas (category 2, *text organisation*).

Based on this work, the marking rubric was refined by removing all semantic overlap. The results from this second series of studies showed that the semantic overlap did to some extent cause artificial consistency in the marking.

Table 2: Extract from the narrative rubric shows semantic overlap of criteria

	Category 1	Category 2
Form of writing	Demonstrates a beginning sense of story structure, for example opening may establish a sense of narrative	Writes a story with a beginning and a complication. Two or more events in sequence. May attempt an ending.
Subject matter	Includes few ideas on conventional subject matter, which may lack internal consistency.	Has some internal consistency of ideas. Narrative is predictable. Ideas are few, may be disjointed and are not elaborated.
Text organisation	Attempts sequencing, although inconsistencies are apparent.	Writes a text with two or more connected ideas. For longer texts, overall coherence is not observable.

Relative crudeness of performance levels

As previously explained, the marking rubric was derived directly from the levels of performance described in the OSF. The explanation that accompanied the introduction of the OSF was that the average student would take approximately 18 months to progress through a level. The levels therefore do not describe and are not expected to describe fine changes in student development.

The statistical analysis of the data provides the opportunity to examine the relationship between levels (as depicted in the marking rubric) and student ability. Figure 2 is taken from the analysis of the writing data and shows that, within a wide ability range, a student would have a high probability of being scored similarly on each criterion. For example, students within the ability range of -3 to +1 logits would have a high probability of scoring all 3s, whereas students in the ability range of +1 to +6 logits would have a high probability of scoring all 4s. Based on the mean scores of students of different age levels, these ability ranges equate to approximately two years of schooling.

Although the marking rubric contained many criteria, and therefore many score points, it provided only relatively few thresholds, or points of discrimination. Essentially, all the information about student performance was obtained from the overall judgment – that is the on-balance judgment of the student's level. All other judgments were replications of that judgment.

Over and above the issues related to the halo effect and the semantic overlap, the marking rubric did not capture the fine changes that can be observed in student writing development. Although there were qualitative differences between the students' written performances, the markers could classify the students only into three or four relatively crude groupings.

Devising a rubric that provides greater precision of student development in narrative writing

Based on an analysis of our findings, it was hypothesised that the general level of description in the framework of how student learning develops did not provide the level of detail we needed for a marking rubric of students' narrative writing. The framework makes no mention of character and setting for example, nor does it articulate in fine detail how students' *sentence level punctuation or punctuation within sentences* develops.

This hypothesis was tested by developing a rubric that captured finer gradations in performance. The new rubric emerged from a close scrutiny of approximately 100 exemplars. We compared the exemplars, trying to determine whether or not there were qualitative differences between them and trying to articulate the differences that we observed. We had

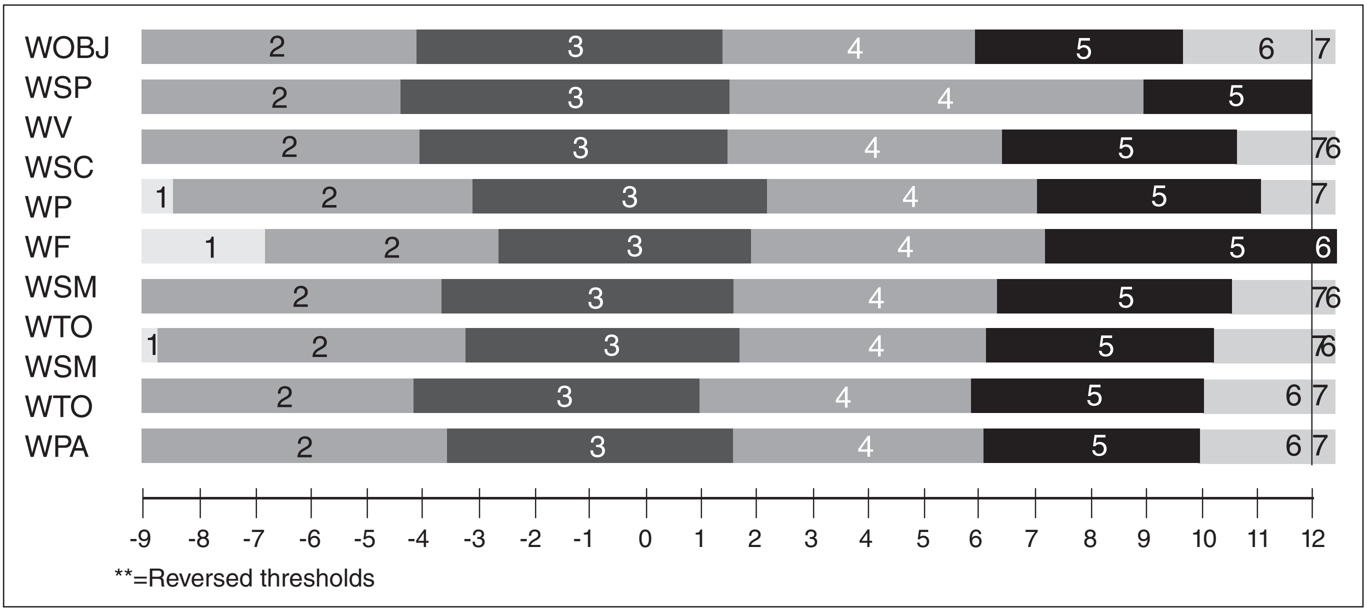

Figure 2: Threshold map showing the relationship between ability and the probability of a score for each criterion.

Table 3: Revised classification scheme for the assessment of writing

Aspect	Score Range	Aspect	Score Range
On-balance judgment	0 – 6	Punctuation within sentences	0 – 3
Spelling	0 – 9	Narrative form	0 – 4
Vocabulary	0 – 6	Paragraphing	0 – 2
Sentence structure	0 – 6	Character and setting	0 – 3
Punctuation of sentences	0 – 2	Ideas	0 – 5
		Total score range	0 – 46

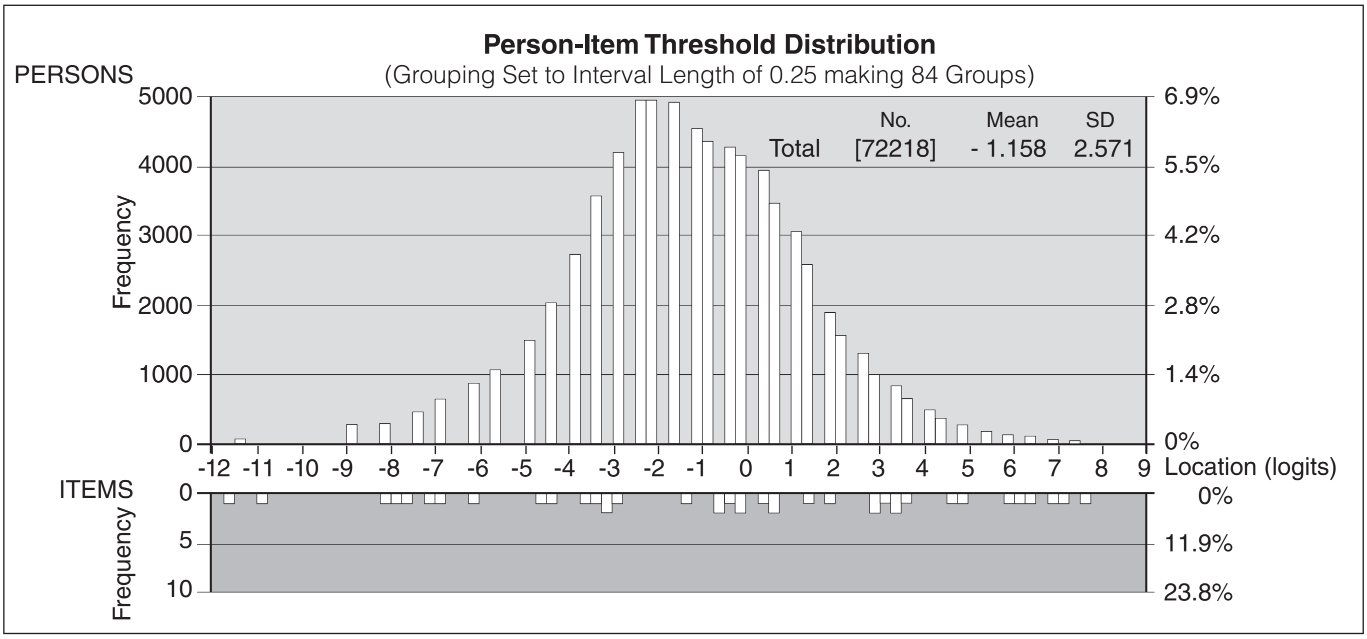

Figure 3: Distribution of students in relation to the thresholds provided in the new rubric

no preconceived notion of how many qualitative differences there would be for each criterion, or that there would necessarily be the same number of qualitative differences for all criteria. Thus the number of categories for each criterion varied depending on the number of qualitative differences we could discern.

For example, in *vocabulary* and *sentence* structure there are seven categories because in a representative range of student performances from Years 3 to 7, seven qualitative differences could be distinguished and described. In *paragraphing* however, only three qualitative differences could be distinguished so there are only three categories. Table 3 shows this revised classification scheme.

The person/item distribution (Figure 3) generated from marking with the new rubric provides greater precision of student development in narrative writing.

Conclusion

Do rubrics help guide and inform teaching practice? Based on this research, the answer to the question on one level is that it depends on the nature of the rubric. In the presentation, a comparison between the criteria in the original rubric with the criteria in the new rubric will be made to illustrate this point. On another level however, this comparison raises questions about the relationship between assessment and teaching, and whether rubrics are sufficient for informing teaching practice.

References

Dimmock, C (2000) *Designing the Learning Centred School: A Cross-cultural Perspective.* Falmer Press, London/ New York.

Heldsinger, S.A. (2009). (In press). Using a measurement paradigm to guide classroom assessment practices in Webber, C.F., & Lupart, J. (Eds.), *Leading student assessment: Trends and opportunities.* Dordrecht, The Netherlands: Springer.

Humphry, S.M., & Heldsinger, S.A. (2009). *Experimental elimination of the halo effect in a performance assessment.* Submitted for publication.

Taggart, G. L., & Wood, M. (1998). Rubrics: A cross-curricular approach to assessment. In G. L. Taggart, S. J. Phifer, J. A. Nixon, & M. Wood (Eds.), *Rubrics: A handbook for construction and use* (pp. 57–74). Lancaster, Pennsylvania: Technomic Publishing Co., Inc.

Popham, W.J. (1997). What's wrong – and what's right – with rubrics. *Educational Leadership,* 55, 72–75.

Wilkinson, A., Barnsley, G., Hana, P. & Swan, M. (1980). *Assessing language development.* Oxford: Oxford University Press.