Assessing Multiple Objectives with a Single Task in Statistics

Jane Skalicky[1]
University of Tasmania
<Jane.Skalicky@utas.edu.au>

In an environment where cross-curricular or interdisciplinary activities are being recommended for classrooms, the question of the assessment of multi-faceted objectives becomes important. Should multiple tasks be employed for multiple objectives or can a single task be devised in a manner such that more than one rubric can be used to assess the multiple objectives? As a first step in the direction of using a single task for more than one objective, a task from statistics is employed to explore the possibility of assessing the two ideas of Expectation and Variation using two rubrics on the same task. Advantages, disadvantages, and educational implications are explored.

The wider context within which this study took place is a reform-based, values-focused curriculum that encourages transdisciplinary activities. The motivation for this study is the reform-based curriculum in Tasmania. The curriculum is centred around five Essential Learnings: thinking, communicating, personal futures, social responsibility, and world futures (Department *of* Education, Tasmania, 2002). Other Australian states and territories are also reconceptualising curricula in terms of similar over-riding big ideas. In the United States, reforms are occurring within the mathematics curriculum itself (National Council of Teachers of Mathematics, 2000) and place importance on the connections among, and applications of, the domains within mathematics. These changes are consistent with the wider reforms.

Changes in curriculum and accompanying pedagogy need to be matched with appropriate changes in assessment practices (Resnick & Resnick, 1985). This poses considerable challenges for the devising of assessment tasks in a transdisciplinary framework (Skalicky, 2004). As activities become more complex, the need to assess many aspects of the expected outcomes becomes important. To have a different task for each area of interest, however, may create a plethora of instruments or questions for students to answer: “efforts to assess thinking and problem-solving abilities by identifying separate components of those abilities and testing them independently will interfere with effectively teaching such abilities” (Resnick & Resnick, 1991, p. 43).

In examining assessment issues within the context of reform-based curricula, it is necessary to look further at the issues surrounding the assessment of multiple objectives. “The assessment of multiple solutions or multiple paths to a single solution, will occur only when we have an approach to assessment that has the same principles as contemporary approaches to mathematics education” (Van den Heuvel-Panhuizen & Becker, 2003). These principles involve valuing the process of student thinking about a problem, not just the solution to it. In both teaching and assessment this requires acknowledging the stages involved in understanding and addressing a task. The rubrics that are devised to judge levels of progress during assessment need to reflect the development taking place and are also likely to be useful in following students’ progress during classroom activities.

A move toward assessment tasks that require students to engage in solving contextually-rich problems addresses some of the issues surrounding assessment of

[1] The participation of Dr Jane Watson of University of Tasmania in the writing of this paper is acknowledged.

mathematical understandings. Such tasks require students to draw on multiple knowledge constructions and sometimes other in-school and out-of-school experiences. Students may privilege some constructions over others, negotiate between areas of knowledge construction, and at the highest level, integrate multiple constructions to fit the assessment task (Kastberg & D'Ambrosio, 2004).

As a starting point for developing multiple rubrics to be used with single tasks, a task from statistics education is considered where multiple objectives are a feature of the expected learning outcomes. Although many topics, such as data collection, graphing, averages, probability, and inference are mentioned in school statistics curricula, recent research has shown that fundamental juxtapositions, for example between signal and noise (Konold & Pollatsek, 2002) or between expectation and variation (Watson & Kelly, 2004), describe the dilemma faced by students when considering statistical problems. Expectation implies pattern and specific values, perhaps based on theory, whereas variation implies change or difference, usually from the pattern or specific value. Traditionally the curriculum has concentrated on the former and often students do not recognise the latter. Tasks are often stated in terms of expectation but given Moore's (1990) view of the omnipresence of variation it is important to follow the development of intuitive ideas about variation when decisions are made about expectation. Although statistics educators are beginning to advocate classroom activities that focus on variation (e.g., Watson, 2002), it is when assessment also addresses variation that teachers and students will take such suggestions seriously.

This study hence focuses on a task that allows for the consideration of variation/noise as well as expectation/signal, using separate rubrics to assess appreciation of the two aspects. The study follows the work of Watson and Chick (2004) who also used two rubrics to assess a single task. In their case the task asked students to describe "unusual" features of a series of three graphs, which had been presented in the media with small errors of representation. Of interest in that study was the notice taken of variation in looking for unusual features as well as students' ability to find the errors in the graphs. They developed two rubrics, one for critical statistical literacy (in finding errors) and one for observing variation. Although Watson and Chick were interested in variation as in the current study, their other rubric related to the specific issue of critical questioning of media sources, an issue for statistical literacy; they did not relate their rubrics to the wider issues of assessment or consider the association of outcomes based on the two rubrics.

The research objectives for this study are based on the feasibility of developing rubrics for a single task that explores aspects of both Expectation and Variation when statistical questions are asked. The feasibility is judged by:

i. the ability to devise meaningful hierarchical rubrics that separate student performance, and
ii. the association of the measures developed to determine if indeed different features are being observed.

Methodology

Task. The task used in this study is presented in Figure 1. The question in the task is stated in terms of Expectation, but most statisticians would believe that solutions should include consideration of the Variation that could influence the expected outcome and hence the confidence with which conclusions are reached. Acknowledgement of uncertainty in predictions made is at the heart of good statistical decision-making.

Two schools are comparing some classes to see which is better at spelling. Look at the scores of all students in each class, and then decide. Did the two classes score equally well, or did one of the classes score better? Explain how you decided. [Two comparisons are shown: Yellow and Brown; Pink and Black.]

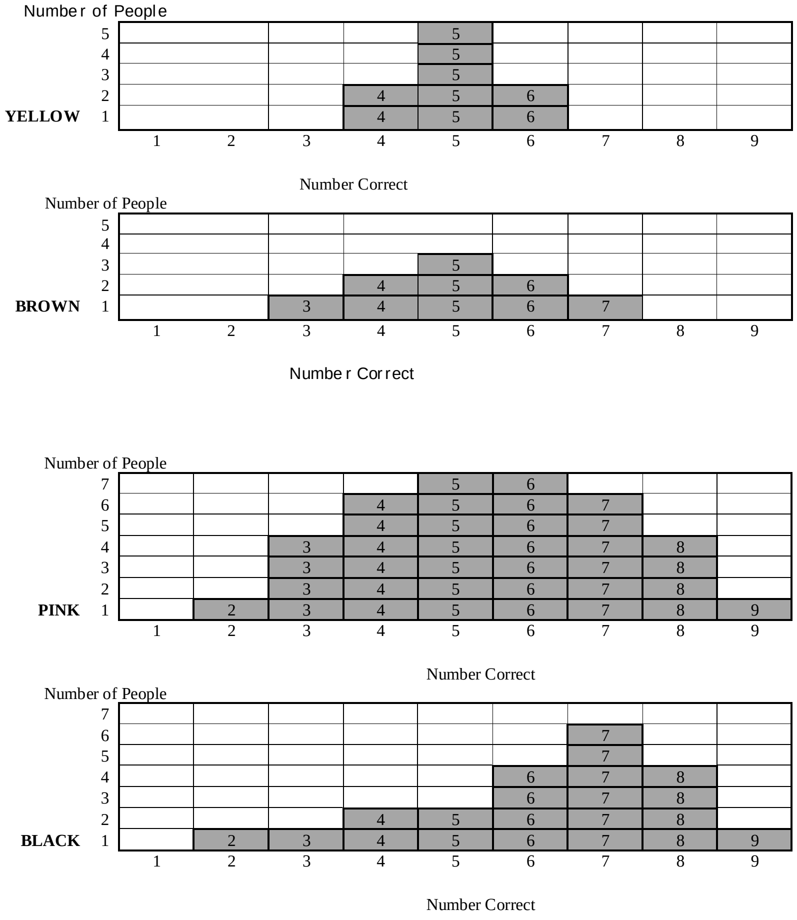

Figure 1. Comparing group task (Watson & Moritz, 1999).

Table 1
Number of Students in each Grade and Year

Study	*Grade*	3	5	6	7	8	9	10	12	13	*Total*
1993/4	Tasmania/SAustralia	23	8	21	8	0	28	0	0	0	88
1997	Longitudinal Tas/SA	0	0	5	8	6	0	15	5	4	43
2000	Tasmania	18	18	0	15	0	15	0	0	0	66
Total		41	26	26	31	6	43	15	5	4	197

Sample. The data were chosen for this study as a representative sample across grades to give the largest possible range of responses, not in order to make predictions about performance across grades. Table 1 contains a summary of the grades and time frame from which the 197 students in the study were interviewed. The responses for 58 students from Tasmania and 30 in South Australia in 1993-94 were analysed only with respect to Expectation by Watson and Moritz (1999), as were 42 of the students from the

Longitudinal data set in 1997 (Watson, 2001). All students in 1997 except one had been interviewed earlier but no students indicated recognition of having answered the protocol before. Data in 2000 were collected in relation to a different project and have not been reported previously; data collection is described in Watson and Kelly (2004). None of these data have been previously analysed with respect to a formal Variation rubric.

Rubrics and coding. The two scoring rubrics and their code levels for the Comparing Groups task are given in Table 2. The Expectation rubric was revised from that devised by Watson and Moritz (1999) in relation to making a decision about differences in performance for pairs of classes on a spelling test. Only the two complex comparison tasks are shown in Figure 1; others involved pairs of sets with six data points in each. The steps in the rubric reflected the structure of responses as suggested by Biggs and Collis (1982; Biggs, 1992), as well as the statistical appropriateness of responses in terms of considering pairs of data sets of equal or unequal size. The first three levels of the Expectation rubric represented consideration of the equal-sized groups, whereas the last two recognised success with unequal-sized groups, based either on proportional reasoning from the graphs or the use of the arithmetic mean. The Variation rubric was devised based on the analysis presented in Watson (2001) where Watson considered how responses employed used the variation displayed in the data as a basis for decision-making. The levels in the Variation rubric again reflected increased structure in the consideration of variation in the pairs of graphs. "More" for example represented a low level justification for the variation between the Pink and Black classes when the Pink class was chosen as better. High level global responses integrated comments about groups of columns with observations about the Black group having relatively more students with higher scores whereas the Pink group had more with middle scores. Using the mean for example was not considered in terms of Variation unless discussed in conjunction with other descriptions of variation.

Table 2
Rubrics for Comparing Groups Task

Level	*Rubric for Expectation*
0	No focus on specific features.
1	Single features of the graph used in simple *equal* size group comparisons.
2	Multiple step visual comparisons or numerical calculations performed in sequence on absolute values for simple *equal* size group comparisons.
3	All available information integrated for a complete response for simple group comparisons. Appropriate conclusions restricted to comparing groups of *equal* size.
4	Multiple step visual comparisons **or** numerical calculations (mean) performed in sequence on a proportional basis to compare groups; **OR** Single visual comparisons used appropriately in comparing groups of *unequal* size.
5	All available information from visual comparisons and calculation of means integrated to support a response in comparing groups of *unequal* size.
Level	*Rubric for Variation*
0	No acknowledgement of variation.
1	Single features evident in only one part of the task. Single column(s) considered: less than or equal to two, or no synthesis; **OR** Observation of 'more' with no further justification.
2	Single features evident in at least two parts. Single column(s) considered: less than or equal to two, or no synthesis; **OR** 'More' with no further justification.
3	Multiple features evident in at least one part. More than two columns considered (but only columns); **OR** Multiple features considered: global plus columns, sequential analysis.
4	Global focus evident in at least one part. Multiple features compared and contrasted.

Results

The distributions of outcomes across the levels of both rubrics for the Comparing Groups task are shown in the row and column totals in Table 3. For this task, the distribution of response across levels for Expectation shows a concentration at Level 2, indicating that half of the students could make multiple-step comparisons and could distinguish differences in groups of equal size. The existence of only one Level 0 response reflects the interview setting where students were encouraged to answer the task and the fact that all students were in at least grade 3 and had seen some type of graphical representation. Only 22% of students responded at Levels 4 or 5, being able to consider groups of different size, using some type of proportional reasoning. The distribution for levels of appreciation of Variation was more uniform than that for Expectation with 28% of responses mentioning a single feature of the graphs related to variation in at least two parts of the task, and 33% mentioning multiple features of graphs, usually related to columns. Only 11% of responses mentioned global features of the graphs in conjunction with multiple features.

Table 3
Association of Levels of Expectation and Variation for Comparing Groups task

Expectation *Variation*	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5	*Total*
Level 0	0	2	19	1	2	0	24
Level 1	1	2	21	4	3	0	31
Level 2	0	9	33	12	1	0	55
Level 3	0	6	25	16	14	4	65
Level 4	0	0	1	2	9	10	22
Total	1	19	99	35	29	14	197

As can be seen in Table 3, the association of response levels for the two rubrics is not strong. An “indicative” correlation between levels (if they are assumed to be equally spaced) is .457, which would explain only 23% of the variance. It appears reasonable to claim that somewhat different aspects of responses to the tasks are being explored with the two rubrics. There are some students who can solve problems involving unequal sized groups without considering more than single aspects of variation and some students who consider multiple aspects of variation but cannot solve problems of unequal sized groups.

Another issue in developing meaningful rubrics is whether they cater for a range of students. The data set in this study allows for the consideration of this by looking at the distributions of response levels across grades. These are shown in Table 4, where it is seen that although there is some indication of overall student improvement in level of performance with respect to Variation and Expectation with grade, what is apparent is that a wide range of performance is catered for and shown at most grade levels. As well, at each level of the rubrics there is a range of performance across grades.

The responses given here are chosen to illustrate the differences in levels of performance for the two concepts. A grade 5 student (Level 1 for Expectation and Level 3 for Variation) considered several aspects of variation but only considered single features in making decisions about expectation.

> Yellow class did better because more people got 5 right. […] The Pink class. [*Why?*] Because more people got 5 or 6 right and then in the Black class most of them got 7 but it is around average that like only one person, but 3 persons got 3 right and so on. [*This means 1 person got 3 and 1 person got 2*] Yes, I know. [*Right. Does it make any difference that there are more people in the Pink class*

> *than in the Black class?*] Yes there’s more people but it doesn’t really make a difference if you have got good spellers in your class, you find that it is different.

Table 4
Response Levels across Grades for Variation and Expectation

Grade	*Response Levels - Variation*						*Response Levels - Expectation*						
	0	1	2	3	4	*Total*	0	1	2	3	4	5	*Total*
3	13	9	12	7	0	41	0	6	32	3	0	0	41
5	4	4	10	8	0	26	1	4	18	3	0	0	26
6	1	2	11	8	4	26	0	4	10	8	4	0	26
7	3	4	7	12	5	31	0	0	13	9	7	2	31
8	0	2	2	2	0	6	0	0	3	1	2	0	6
9	2	9	10	21	1	43	0	4	17	8	11	3	43
10	1	1	2	6	5	15	0	1	5	2	3	4	15
12	0	0	1	0	4	5	0	0	1	1	0	3	5
13	0	0	0	1	3	4	0	0	0	0	2	2	4
Total	24	31	55	65	22	197	1	19	99	35	29	14	197

A grade 12 student (Level 2 for Expectation and Level 4 for Variation) compared and contrasted the features of the two graphs in terms of variation for each pair but could not deal with unequal sized groups or resolve the expectation issue with equal sized groups.

> The people in the Brown class, like one person actually scored a 7, both classes have two 6’s, I think having more people with a score of 5 is actually better than having one person with a score of 7 … the Yellow class because they have got more people scoring an average score and this class only has a few people and each side is sort of scattered around and this class is quite direct and has quite a few people getting an average mark. […] Both the Black and the Pink class their 9, 8, 7 is equal but in the Pink class their 5 and 6 is very high compared to the Black class so you would think they would be better people in that class, because … there are quite a few people in that class but that is equal to the Pink class, like you can see from that those ones there are exactly equal where as those ones there have suddenly dropped, where as here the marks have come up quite high like there are quite a few people getting those scores, so I think the Pink class would have to be smarter.

In contrast, a grade 9 student (Level 4 for Expectation and Level 0 for Variation) immediately focused on a concise mathematical calculation to solve for the mean of each class and did not consider variation.

> [Student writes totals under each of the Black columns and totals for Pink columns] [*Interviewer gives student the totals for each – 130 (Black); 198 (Pink)*] [Student uses calculator a few times: writes numbers at top of each column to get total number of students in each class; writes down “Ave. 5.5” (Pink); “Ave. 6.19” (Black) and writes down: “Black class scored better because the Average score was .69 better” (uses calculator once here)] .

There were 10 students who performed well on both Expectation and Variation. A grade 13 student (Level 5 for Expectation and Level 4 for Variation) used both proportional reasoning and appropriate descriptions of variation.

> They both had the same average of 5 … so they sort of scored equally ... but the Brown class had a greater variation in scores, whereas the Yellow had a greater concentration. […] Well this time there are obviously more people in the Pink class than in the black class, but I’d say that the Black class scored better in that the average is higher because there are a greater proportion of them in the 6, 7, 8 range and whereas the Pink class had in the 5 and 6 range. And so the Black class had a higher average. [*You’d use average for all of the pairs?*] Yes.

A small number of students scored low on both Expectation and Variation. A grade 5 student (Level 0 for Expectation and Level 1 for Variation) showed little appreciation of the task except for noticing "more" and "most" in the graphs.

> Probably the Yellow … because more people got a … one person got it all right … or something. […] Pink. [*Why?*] Because they Pink class they all were … most of them did well. But on the Black class they didn't do too bad but … [*Not quite so well?*] Yes.

About a quarter of the students interviewed fell within the middle levels for Expectation and Variation and were able to consider multiple aspects of both Expectation and Variation within this one task. A grade 3 student (Level 3 for Expectation and Level 2 for Variation) could handle equal sized groups, for example by balancing the Yellow and Brown classes but not unequal groups, focusing on "more".

> About the same. [W*hy*?] Because they [Yellow] have got more 5's, two more, and these [Brown] have got an extra 7 and 3. […] Pink. [W*hy?*] Because they have got more shaded and … [*Why do you think that is?*] Because they had more students. [*Right, does it matter?*] It depends [shrugs shoulders].

Discussion and Implications

In terms of the research objectives of this study, judgment of the meaningfulness of the hierarchical rubrics rests to some extent with teachers or measurement specialists who would use them for specific assessment purposes. The fact that they can be justified by increasing mathematical appropriateness and structure in the case of Expectation and by increasing recognition and structural complexity in the case of Variation, supports the validity of their hierarchical meaning. It also demonstrates the value of assessing both constructs within the one task and begins a necessary consideration of construct validity (Messick, 1994) when assessing multiple objectives in a single task, which may provide helpful insights in a move toward employing more complex tasks in reform-based curricula**.** The spread of responses for students from grade 3 also suggests appropriate coverage for school students. The relatively low percent of shared variance supports the view that somewhat different aspects of statistical understanding are being tapped.

The results suggest that a single task in statistics can be used to provide information about students' understanding of both Expectation and Variation. The hierarchical rubrics used to assess student performance on the task provide meaningful feedback on the knowledge and thinking of students on multiple dimensions and this has implications for the teaching of Expectation and Variation. What one would like to achieve by the end of secondary schooling is the highest level of understanding on both Expectation and Variation. It would appear, however, based on students currently in the school system in Australia that development does not occur exactly in parallel for the two aspects of understanding. If teachers are aware of this they can adapt classroom discussion to cater for consideration of the topics in conjunction with each other. Further if assessment practices begin to include recognition of variation, it may achieve a higher profile in classroom activities.

The advantages of using two rubrics for a single task have been covered in the preceding paragraphs. Are there disadvantages? It might be claimed that the tasks should somehow be phrased to alert students to the need to discuss variation. It could be argued that the presentation of graphs should be a cue to consider the variation presented in the distributions. Given Moore's (1990) views on the importance of variation as the foundation of statistical investigation, it seems reasonable to assert that it should not be necessary to make a request for its discussion – it should be a natural part of the consideration of every problem encountered.

As reform in mathematics education values the processes of student thinking and the connectedness of topics, assessment needs to provide teachers with relevant and useful information that will inform the teaching process from as many perspectives as possible. With regard to broader curriculum reforms, single tasks that can be used to assess multiple knowledge constructions, as suggested in this study, may provide not only teachers, but also writers of external assessment items, with a framework from which to begin to match assessment more closely with the teaching and learning needs of students.

Acknowledgement. This research was supported by Australian Research Council Grant. No. DP0208607. Jonathan Moritz and Ben Kelly assisted in coding the data.

References

Biggs, J.B. (1992). Modes of learning, forms of knowing, and ways of schooling. In A. Demetriou, M. shayer & A. Efklides (Eds.), *Neo-piagetian theories of cognitive development: Implications and applications of education* (pp. 31-51). London: Routledge.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Department *of* Education Tasmania. (2002). *Essential learnings framework 1*. Hobart: Author.

Kastberg, S.E., & D'Ambrosio, B.S. (2004). *The role of contextually-rich items in assessing student learning.* Paper presented at the International Congress on Mathematics Education, Copenhagen, Denmark. Available at http://www.icme-organisers.edk/tsg27/papers/12_signe_et_al_fullpaper.pdf

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*, 259-289.

Messick, S. (1994). The interplay of evidence and consequences in validation of performance assessments. Educational *Researcher, 23* (2), 13-23.

Moore, D.S. (1990). Uncertainty. In L.A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

National Council of Teachers of Mathematics. (2000) *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Resnick, D.P., & Resnick, L.B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher*, *14*(4), 5-21.

Resnick, D.P., & Resnick, L.B. (1991). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (p. 37-75). Boston: Kluwer.

Skalicky, J. (2004, December). *Quantitative literacy in a reform-based curriculum and implications for assessment.* Paper presented the Australian Association for Research in Education Conference, Melbourne. Available at http//www.aare.edu.au/04pap/alpha04.htm

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105-110.

Van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. In A.J. Bishop, M.A. Clements, C. Keitel, J. Kilpatrick, & F.K.S. Leung (Eds.). *Second international handbook of mathematics education* (Pt. 2, pp. 689-716). Dordrecht: Kluwer.

Watson, J.M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics, 47*, 337-372.

Watson, J.M. (2002). Lessons from variation research II: For the classroom. In M. Goos & T. Spencer (Eds.), *Mathematics - making waves*. (Proceedings of the 19th Biennial Conference of the Australian Association of Mathematics Teachers Inc., Brisbane, pp. 424-432). Adelaide, SA: AAMT, Inc.

Watson, J.M., & Chick, H.L. (2004). What is unusual? The case of a media graph. In M. Johnsen-Høines & A. B. Fuglestad (Eds.), *Proceedings of the 28th annual conference of the International Study Group for the Psychology of Mathematics Education* (Vol. 2, pp. 207-214). Bergen, Norway: PME.

Watson, J.M., & Kelly, B.A. (2004). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics and Technology Education*, *4*, 371-396.

Watson, J.M., & Moritz, J.B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37,* 145-168.