From normal curve to slippery slope

John H Court PhD.

Australia,(Retired)

Published in *Testing International*, 19 July 2008, p12-13

In 1997 Boyer wrote **"**We urge that student evaluation be used in making decisions about tenure and promotion. But for this to work, procedures must be well designed and students must be well prepared." (p.40)

In the nineties, a national initiative across Australian universities resulted in development of the Graduate Course Evaluation Questionnaire (GCEQ) to determine student satisfaction following graduation and to benchmark all participating universities. The 25-item instrument, in paper and pencil format was well researched for psychometric properties. So far, so good.

Taking up the challenge to improve student learning experiences, one university responded by introducing in-house evaluations for all undergraduate courses, borrowing from the national instrument. A series of steps over a five year period indicates how easy it is to move from a well constructed instrument with a well-defined purpose, to the accumulation of numbers with no credibility, used for different purposes and in different ways, yet leading to significant decisions way beyond the original purpose. An interview with the academic responsible for these developments is revealing, and brief excerpts are included below.

First the GCEQ was shortened from 25 items to 10 by a committee. "No study has shown that this selection correlates with the larger item pool, and we don't know the validity of the currently used instruments."

The paper and pencil administration in class was followed initially, but gradually changed across to requests for staff to encourage to online responding. The rationale for this was "to maximize anonymity and because we can't afford paper and pencil administration".

This major shift calls into question any residual validity or reliability. It means that the instrument is undertaken at various times around the end of a course (invariably before final assessment is completed, though questions relate to that), and the percentage of students deciding to respond could b expected to plummet. In fact comparisons across several courses were possible, showing an average response rate for paper and pencil of 59%, with online averaging 11%. Recognising the problems, a researcher spent a year trying to enhance response levels and reported an increase of 17%, which actually meant a shift from 11 % to nearly 13%. This might suggest there is an inherent problem in the approach.

Under such circumstances all assumptions about a normal distribution have to be discarded. Academics recognize that this small number is largely composed of the disaffected and the very enthusiastic in classes, so this bimodal distribution is of little help in identifying good learning, or in recommending changes.

The shift of modes raises an ethical dilemma which has been addressed by Susan Whiston (2000) viz. "If the instrument is an adaptation of a paper and pencil instrument, then the evaluation of the psychometric qualities must include an analysis of the equivalency of the two forms of the instrument." (p.352)

Nonetheless the data have continued to be a source for evaluating course outcomes . In addition, they have moved to become compulsory for all courses at all levels, forming a component of evaluations for academic promotion.

After five years of usage, do we have evidence of improvements in teaching arising from these data? "No". Do the undergraduate data generally conform to the normal curve of distribution? "We haven't looked at that".

A psychologist considering the student experience might ask what is going to happen if students are invited at the end of every course in their degrees to respond to an online satisfaction survey. Since the responses are anonymous there is no capacity to offer reinforcement for responding, so one could predict decreasing interest through the years of the first degree. Students continuing on to higher levels may encounter 30 or more such evaluations. Not surprisingly therefore, with contingent reinforcement absent, in graduate programs response numbers are often in single figures, heading for extinction of the response. This is a pity since pre-existing paper and pencil evaluations ran at better than 85%.

Remarkably, therefore, the next step was to make these data available as essential information for academics seeking promotion. It would seem to follow that high ratings might be viewed favourably. However, responding to the comment "Presumably data from high quality instructors is negatively skewed" the response was "They should be but we don't know".

These steps away from use of psychometrics to an ad hoc application of numbers are not presented as a criticism of the particular location where these decisions were made. They are identified as an example of what can happen all too easily where a results-based institution is looking for ways to demonstrate its attention to accountability. This often links to funding decisions as well as perceived reputation, so the pressures for numbers are great. So are the hazards of adapting, abbreviating, and modifying procedures without developing evidence on the effects such changes have on response patterns.

There are times when a misguided commitment to test data that appear to have a respectable pedigree can lead to a situation where bad data are worse than no data. Psychologists required to participate in such procedures face interesting ethical dilemmas, poised between their code of professional conduct and the expectations of employers.

Whiston, S. C. (2000). *Principles and applications of assessment in counseling.* Brooks/Cole Thomson Learning.

Boyer, Ernest L. (1997). *Scholarship reconsidered: Evaluation of the professoriate*. San Francisco: Jossey-Bass.