Erosion of Peer Review?

W e respectfully disagree with the Editor’s opinion on the consequences of the changes in the proposal review process for the 2014 NCME Annual Meeting expressed in one of his recent editorials (Briggs, 2014).

The core of the Editor’s argument can be summarized follows: The numbers of submission for the 2014 meeting went up considerably, the same happened to the number of accepted proposals, thus the quality of the average NCME presentation has likely decreased in quality. The sizes of both changes seem to reinforce the argument. The number of submissions went from 554 for the 2013 Annual Meeting to 772 for 2104; at the same time the number of accepted proposals increased from 238 to 464.

The Editor’s second premise reminds us of the well-known Taylor–Russell (1939) tables for the success of prediction in personnel selection: If in a series of selection decisions the selection ratio goes up, then—all other things equal—the success ratio goes down. For their logic to be valid, the all-other-things-equal clause is critical, though. The Taylor–Russell tables hold only under the assumption that the validity of the selection instrument—the correlation between the predictor and test scores in the bivariate normal distribution underlying them—remains constant. And, we believe that it is exactly the validity of the review process that was impacted dramatically by the changes introduced for the 2014 Annual Meeting.

For the 2013 Annual Meeting, 407 volunteers evaluated 554 submitted proposals. With this massive number of volunteers, it is a nearly impossible challenge to match their expertise with the topics of the proposals. All volunteers worked independently in sets of three per proposal. If one of them gave less than perfect rates on some of the criteria, given the low number of slots for presentations available the proposal was most likely rejected. No wonder that, for the last several years, the program chairs had to deal with large numbers of complaints when their acceptance/rejection decisions were released.

For the 2014 meeting, the goal was to increase the number of slots and at the same time improve the validity of the review process. This is what was introduced: The program chairs assigned all proposals to one of ten thematic “bins” of roughly the same size. These relatively homogenous bins were chosen following an analysis of the program books for the last several years. For each topical bin, the program chairs selected a panel of three senior experts. Each panel had to classify the proposals in its bin into three categories (definitely accept; accept if possible; definitely reject) and then rank the ones in the middle category. An essential feature of the review process was that all three panel members had to reach a consensus on their classifications and rankings before they were submitted to the program chairs. Except for “rounding error,” the program chairs filled the slots in their program beginning with the highest category and continued with the ranked proposals in the middle category.

We believe that senior reviewers selected by program chairs on the basis of their expertise produce more valid selection decisions than an extremely large set of self-selected reviewers. Also, the necessity to reach consensus seems a more valid way to deal with outlying ratings than just averaging them. If Educational Measurement: Issues and Practice would have had its manuscripts for the current volume reviewed by 407 self-selected volunteers instead of 30 appointed reviewers, it might already have taken complaints. There may have been some unplanned variations in the ways the panels operated, leaving room for further improvements this year, but for the first time since several years, the program chairs did not receive any complaints about their rejection decisions.

Ironically, the higher validity of the new review system follows from the same logic of the Taylor–Russell tables as invoked by the Editor. The new review system reduced the selection ratio for the reviewers relative to those for 2013 by a factor of 13.6 (30 instead of 407)—a reduction much more dramatic than the increase in acceptance rate for the proposals by the factor of 1.9 (464 instead of 238) reflected on by the Editor.

In fact, we could even invoke the same logic one more time: Each of the meeting attendees had 1.9 times as many presentations to select from for the same number of days. We hope that, as the result of their smaller selection ratio, they were able to find more highly relevant presentations and had an enjoyable experience.

While we seriously disagree with the editor’s assessment of program quality, we do thank him for a number of specific suggestions for further improvements to the process. We are intending to bring back discussants for the paper presentations, increase the number of panels somewhat, and work to standardize rating scales used in assessing the quality of the proposed presentations in putting together the 2015 program.

Wim van der Linden and Lauress Wise

References
