#3F7 (27306)
Four years of MCQ results reveal consistent patterns in certainty and safety, insightfulness and foresightfulness

Mike Tweed*, University of Otago Wellington, Department of Medicine, Wellington, New Zealand
Tim Wilkinson, University of Otago Christchurch, Medical Education Unit, Christchurch, New Zealand
Sarah Stein, University of Otago, Higher Education Development Centre, Dunedin, New Zealand
Jeff Smith, University of Otago, College of Education, Dunedin, New Zealand

Background: Insightfulness and foresightfulness can be inferred from responses to MCQs as, respectively the certainty in correctness and avoidance of options that might be unsafe, certainty safety response questions (CSRQs). This investigation is an analysis of lower stakes summative assessment results found over several years' cohorts.
Summary of Work: CSRQs were used as part of the assessment package for 329 Year 5 medical students over 4 years, 2011-2014. Incorrect responses had been classified by safeness by an expert panel. Analyses compared response certainty, safety and student performance groupings overall and across years.
Summary of Results: Comparing 2011-12 with 2013-14, there was no difference in proportions of responses (correct, unsafe, unsafe of incorrect), for performance groups (low scorers, below standard scorers, above standard scorers, excellent scorers) and certainty (low, moderate, high). As certainty in responses increased, the mean proportion of all responses correct increased and the mean proportion incorrect and unsafe decreased. Similar patterns were seen for the whole cohorts and performance groupings. As certainty in responses increased, the proportion of incorrect responses that were unsafe increased. This pattern was only seen for the above standard groupings.
Discussion and Conclusions: Across four years all student groups, including those below standard, demonstrate insightfulness, with increasing correctness with certainty. Although those scoring above standard have insightfulness, they have less foresightfulness with more unsafe decisions, when incorrect, as their certainty increases.
Take-home messages: CSRQ format provides consistent results on cohort insightfulness and foresightfulness. Lower performers do have insightfulness. Those just above standard have limited foresightfulness.

#3F8 (28263)
Effects of Gaming and Guessing on Script Concordance Test Scores

Stuart Lubarsky*, McGill University, Centre for Medical Education, Canada
Valerie Dory, McGill University, Centre for Medical Education, Canada
Sarkis Meterissian, McGill University, Centre for Medical Education, Canada
Carole Lambert, Universite de Montreal, CPASS, Canada
Robert Gagnon, Universite de Montreal, CPASS, Canada

Background: In an SCT, examinees are asked to judge the effect they think a new piece of clinical information has on a proposed hypothesis. Answers are collected using a Likert scale (ranging from -2 to +2, with '0' indicating no effect), and compared to the answers of a reference panel of 'experts'. It has recently been argued, however, that SCT may be susceptible to the influences of gaming and guesswork. This study aims to address some of the mounting concern over the response process validity of SCT scores.
Summary of Work: Using published datasets from 3 separate SCTs, we computed the score a hypothetical student would obtain on each of the tests if he 1) guessed random answers and 2) deliberately answered '0' on all test items.
Summary of Results: A simulated random guessing strategy led to scores at least 2 SDs below mean scores of actual respondents. A simulated 'all 0' strategy led to scores at least 1 SD above those obtained by random guessing. In one dataset, stepwise exclusion of items with modal panel response of '0' to 10% of the total number of test items yielded hypothetical scores around 2 SDs below mean scores of actual respondents (similar to guessing).
Discussion and Conclusions: Random guessing led to lower scores than those obtained by actual respondents. An 'all 0' strategy, however, demonstrated evidence of artificial score inflation. Our findings pose a significant threat to the SCT's validity argument. 'Testwiseness' is a potential hazard to all testing formats. We found that limiting the proportion of SCT items with a modal panel response of '0' to no more than 10% of the total number of test items mitigated the effects of the 'all 0' strategy.
Take-home messages: Gaming and guesswork pose a legitimate threat to the validity of script concordance testing, but can be mitigated through diligent item construction.