#10E Research Papers: Testing of and for Learning

**Location:** Hall 3d

**#10E1 (75)**

**Testing Test-Enhanced Continuing Education - A Randomized Controlled Trial**

**Mark Feldman, University of Toronto, Hospital for Sick Children, Toronto, Canada**

Oshan Fernando
Michelle Wan
Tina Martimianakis
Kulamakan Kulasegaram

**Introduction:** Enhancing retention and application of knowledge through continuing professional development (CPD) is challenging. Recertification may require CPD using assessment and feedback. Test-enhanced learning (TEL) is effective for increasing durability of teaching in undergraduate medical education. However, the value of TEL in CPD is under-researched. Limited data suggests that TEL for CPD may be inefficient or ineffective. Furthermore, how best to deploy testing and its impact on knowledge is still poorly understood. In this pragmatic RCT we investigated the use of an efficient, pre/post TEL package at a national conference and its impact on knowledge as well as self-reported learning behaviours.

**Methods:** Set in Toronto, Canada, we conducted a randomized controlled trial of TEL for paediatricians who registered for an annual CPD update conference. Participants were randomized to receive 5 standalone pre-workshop MCQ tests followed by the same post-workshop MCQ tests (with feedback) or to receive no pre/post-workshop testing. MCQ tests were delivered electronically before and after attending the conference. Results from the pre-tests were given to workshop teachers to supplement their needs-assessments. Post-tests results and explanations were given to participants. Our primary outcome measure, a clinical vignette based retention and application test, with different MCQs, was delivered to all participants a month following the conference. We analyzed our primary outcome with each workshop considered to be an individual ‘study’ and cumulatively compared performance for the intervention across workshops using a random-effects meta-analysis model. Secondary outcomes examined participant satisfaction, self-reported changes in learning behaviour and completion rates.

**Results:** The study was conducted for 15 individual workshops whose leaders provided MCQs. One-hundred eighty-six participants across Canada consented to participate in the study; each was randomized to receive testing for some workshops and to the control (no testing) group for others. For our primary outcome, one-hundred twenty-six participants completed the retention test. On average, retention test scores were significantly higher for participants (71.2%) who received TEL than those randomized to workshop-only (60%); this was a moderate effect size (Hedge's g of 0.46, 95% CI: 0.26-0.67). The majority of participants (>80%) agreed that pre-tests helped identify knowledge gaps and enhanced learning at the workshop and that post-tests with feedback helped verify learning.

**Discussion and Conclusion:** Our TEL strategy demonstrated benefit for both perceived and objective measures of knowledge retention with moderate effect size. How TEL will transfer to clinical outcomes was beyond the scope of our inquiry. Our results may underestimate the benefit of TEL. Repeated inclusion and crossing over of participants may attenuate effects. Only 5 MCQs were assigned per test. Effect size may increase with MCQ number and quality. Even so, our results indicate that this portable model of testing with feedback can be broadly leveraged to efficiently and effectively improve outcomes of CPD. Evaluating clinicians not only enhances learning but may also help gauge knowledge; requisite for competency-based CPD. These findings will improve practice for CPD in Canada and fill a gap in the CE literature internationally.

**References:** Neurology. 2015 Feb 17;84(7):748-54; J Con Educ Health Prof. 2015 Spring;35(2):119-22

**#10E2 (217)**

**Avoiding medical skill decay: can we apply the testing effect?**

**Dario Cecilio-Fernandes, Center for Education Development and Research in Health Professions (CEDAR) - University of Groningen, University Medical Center Groningen, Groningen, Netherlands**

Fokie Crossen, University of Groningen, Groningen, Netherlands
Stephan Mooibroek
A. D. (Debbie) C. Jaarsma, University Medical Center Groningen, Groningen, Netherlands
René A. Tio, University Medical Center Groningen, Groningen, Netherlands

**Introduction:** The acquired skill will decay after a period of non-use. From cognitive psychology, research has shown that long-term knowledge retention increases more by being tested than by repeating the learning material(1). However, evidence for this so-called testing effect in medical skills training is lacking. In this study, we investigated whether the testing effect would improve students’ long-term retention of making a transthoracic echocardiogram (TTE). Since all skill acquisition starts with knowledge, novices would benefit from retrieving the required knowledge. We hypothesized that an interim knowledge test would improve students’ skills retention, since their performance is still not automatized, which requires students to retrieval their knowledge.

**Methods:** 43 medical students without TTE experience participated in this experiment. In the first session, students took a baseline theory test consisting of 20 questions on structures, functioning of the heart and about manipulation of the probe. Then, students watched a video which explained the basic anatomy and function of the heart, and gave an introduction on echocardiography and the required views. Afterwards, students practiced in groups of two to four people. The students practiced six
views on the simulator, followed by a theory and TTE test. The practice time was equal for all students. After 3 weeks, all students performed a theory and/or TTE test, and were assigned to one of the four groups; group 1 only knowledge test; group 2 only a practical test; group 3 a knowledge and practical test; group 4, the control group no test in between. After eight weeks, all students took a both tests.

Results: There were no significant differences between groups for neither baseline nor end of training measurements. The interim knowledge test had an influence on the follow-up knowledge test (F(1,31) = 1.20, p = .019) whereas the practical interim test not taking both interim tests had a significant effect. The interim practical test had an influence on the follow-up practical test (F(1,31) = 2.99, p = .006) whereas the knowledge interim test not taking both interim tests had a significant effect.

Discussion and Conclusion: We had hypothesized that an interim knowledge test would improve students’ retention on making an echocardiogram. However, our results demonstrated that students only benefited on the follow-up knowledge test. It seems that making an interim practical test would be more beneficial for students in making an echo. This may perhaps be explained by the fact that making an echo is a perceptual-motor tasks in which students learn in an intuitive fashion by only practicing. In conclusion, the testing effect may avoid skill and knowledge decay. In addition, it may be concluded that the testing effect is very specific to the type of knowledge: theoretical vs practical.


#10E3 (231)
Very Short Answer and Single Best Answer Questions: comparison of precision, students’ performance, discrimination and acceptability

Amir H Sam, Medical Education Research Unit, Imperial College London, School of Medicine, UK
Samantha M Field, Medical Education Research Unit, Imperial College London, School of Medicine, UK
Carlos F Collares, Department of Educational Development & Research, Maastricht University, Netherlands
Cees Van De Vleuten, Department of Educational Development & Research, Maastricht University, Netherlands
Val Wass, Faculty of Health, Keele University, UK
Keean Schupke, Fry-it, London, UK

Introduction: Single Best Answer (SBA) questions are used widely in undergraduate and postgraduate assessment. However, there are pedagogical concerns that SBAs may not provide a true reflection of knowledge as they rely on recognition of the answer rather than recall of knowledge. [1] We developed a novel assessment tool that facilitates efficient marking of open-ended ‘Very Short Answer’ (VSA) questions. We compared VSAs with SBAs in terms of precision (reliability), student performance and discrimination. We also evaluated the acceptability of VSAs according to the opinions of the students.

Methods: Medical students sat a 60-question formative examination in either VSA then SBA format (group 1, n=155) or the reverse (group 2, n=144). VSA questions were presented on a tablet with space to type a short answer (1-4 words), which was computer marked. Two examiners reviewed the machine-marked VSAs taking on average 1-36 minutes per question for all students. The SBA assessment was delivered in paper format with answers recorded on a machine-marked scoring card.

Item response theory (IRT) analysis was used to estimate difficulty and discrimination parameters, precision and latent scores for both response formats in the two groups using the measurement alignment approach [2]. Angoff and Ebel methods were used to set the cut scores for both question formats. Positive cueing, defined as percentage of questions answered correctly in the SBAs and incorrectly in the VSAs, was also assessed. A post-test survey was conducted using a 5-point Likert scale.

Results: Mean precision estimates (reliability) were higher in VSAs (0.91 and 0.90) than the SBAs (0.83 and 0.85) and the differences between groups were significant (Games-Howell post-hoc test). Item discrimination was significantly higher for VSAs according to point biserial correlations, but not according to IRT discrimination parameters. Positive cueing was seen in 19% of questions. When students sat the VSA test first (group 1), the mean latent score was lower than when they sat the SBA test (Cohen’s d coefficient = 0.40). The difference in mean latent scores decreased in group 2 (Cohen’s d coefficient = 0.14). A higher proportion of students were below the cut scores in the VSAs. Although 81% percent of students thought the VSAs were more difficult, 69% found them more authentic.

Discussion and Conclusion: The results of this study demonstrate for the first time the utility of VSA questions as a novel assessment instrument. Student performance in the SBA test was significantly better, likely due to cueing: cueing is not a feature of VSA questions. VSA questions also enable testing of knowledge that is known to be difficult to assess in SBA format and avoids the need to identify congruent (but incorrect) alternative options. VSA questions demonstrated higher precision but evidence on item discrimination is heterogeneous. VSA questions are more authentic and, therefore, may have superior consequential validity.

#104E (263)
Can practice make perfect? Learning Curves inform assessment of ECG competency

Rose Hatala, University of British Columbia, Vancouver, Canada
Jacqueline Gutman, New York University School of Medicine, New York, USA
Marc Triola, New York University School of Medicine, New York, USA
Martin Pusic, New York University School of Medicine, New York, USA

Introduction: Learning curves (LCs), produced by sampling a large number of cases and compared against expert performance, are an ideal method of competency-based assessment. Learning curves emphasize the level of competency all learners must achieve, while describing the individual effort required to achieve this goal. Learning to interpret visual information such as an ECG in the clinical setting is often inefficient and ineffective and formal assessment is limited. The advent of large computerized image banks affords access to large volumes of cases using principles of mixed practice, immediate feedback and assessment through LCs.

We describe an on-line ECG learning system using LCs as a competency-based approach to diagnosing a learner, predicting the trajectory of their learning, and suggesting the volume of cases required to attain competence.

Methods: We developed an on-line program of 293 anonymized ECGs collected from an electronic patient database. Each ECG was diagnosed by 3 cardiologists and annotated to highlight key features. ECGs were presented in mixed order, with 114/293 (39%) normal ECGs. Voluntary participants reviewed each blank ECG and selected their final diagnosis. Immediate feedback was provided as the ECG with computerized interpretation, expert annotation, and correct diagnosis.

In 2015, 63 medical students, 10 PGY-1 residents and 5 cardiology fellows each completed at least 60 ECGs. Data included descriptive data and diagnostic accuracy per ECG completed. We calculated sensitivity and specificity and analyzed learning curves by level of expertise. We reported the parameters of a mixed-effects logistic regression on 9054 completed cases where diagnostic accuracy predictors included the log-transformed number of cases already completed, and the participant’s training level. We also report the marginal predicted probability upon completing a 100th case within the system.

Results: Time spent per ECG was 41.55s for students, 68.3s for residents and 14.0s for fellows. Data analyses are presented in the Table. The predictive model showed a statistically significant effect of number of ECGs read (log odds = 18.1, SE = 0.025) and expertise (log odds fellows = 7.38, SE = 0.69), on accurate response. Interestingly, each group learned at approximately the same rate, albeit from different starting y-intercepts.

Discussion and Conclusion: An on-line ECG learning system, based on principles of mixed and deliberate practice and attending to real-world prevalence of normal cases, was effective at improving learners’ competency in ECG interpretation. While individual differences exist, learning curves by level of expertise were established, which allows establishing levels of prior knowledge (y-intercept), prediction of the trajectory of learning (achievement at 100th case) and the volume of cases required to achieve a competency threshold. The method also clearly demonstrates the progressive decrease, with practice, in between-learner variance. These properties were determined without having to resort to separate assessments, but rather while the individuals were practicing and improving their skills.


#1085 (261)
A pluralistic approach to standard setting in knowledge tests

Jonathan Darling, University of Leeds, Leeds, UK
Matt Homer, University of Leeds, Leeds, UK
Rachel Isba, North Manchester General Hospital, Manchester, UK

Introduction: Test-centred approaches to standard setting (such as Angoff or Ebel) for knowledge tests are problematic because expert judges may struggle to predict item difficulty. They are also demanding of assessor time. An alternative approach by Cohen suggests use of high performing students (eg those at the 90th or 95th centile) within the assessment cohort to set the standard for the exam, by fixing the passing score as a proportion of their score. This method is straightforward to run, and is attractive because it gives a ‘student view’ of the difficulty of the exam and adjusts the passmark accordingly. Rasch analysis allows the difficulty of each item to be calibrated against other items, both within one assessment, and across multiple assessments through use of common items.

Methods: We used Rasch analysis to compare Ebel and Cohen approaches to standard setting in a summative Year 4 medicine knowledge test in adjacent years (2014 and
Each test consisted of 200 single best answer items, covering paediatrics, psychiatry, obstetrics and gynaecology, emergency and critical care, and cancer and continuing care. Ebel was the main standard-setting method, where multiple examiners agreed the difficulty and importance of each item (each on a three point scale), and this then determined the passing score according to a pre-agreed matrix. Item-level scores were then summated to give the passing score. For the Cohen approach we calculated the 90th and 95th percentiles of student total scores to compare the difficulty of both tests. Finally we used Rasch analysis to link the two tests, using 50 anchor items shared by both tests, distributed across all the speciality areas.

**Results:** There were 293 and 274 candidates in 2014 and 2015 respectively. 2 and 9 poorly performing items were removed from 2014 and 2015 tests respectively. The Ebel standard in 2014 was 55.6% and 55.0% in 2015, suggesting similar test difficulty. However the Cohen percentiles (2014/15) respectively were 74.2/79.6 (90th) and 76.9/81.7 (95th), equivalent to a difference in passing score (based on the 90th and 95th percentiles of 4.0% and 3.5% respectively, with the 2014 test being harder. Rasch analysis showed good fit to the model, demonstrated that the tests were very close in difficulty, and that the Ebel standard setting (rather than Cohen) had maintained an equivalent standard year-on-year. The two cohorts were shown to be of different average ability (2015 were more able).

**Discussion and Conclusion:** Using Rasch analysis we have compared Ebel and Cohen standard setting methods across two exam cohorts. In this study, Ebel was shown to be effective at maintaining a similar standard across the two cohorts, while the Cohen method’s assumption that bright students will demonstrate equivalent performance in different year groups was shown not to hold, since the 2015 cohort were a stronger group sitting an exam of similar difficulty to 2014. Our research indicates that pragmatic use of multiple approaches in concert is likely to result in a deeper understanding of exam data, and a more nuanced standard setting process. In the last part of this presentation we will discuss how a pluralistic methodological approach would work in practice.