RIME Session: Impact of Assessment

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

Methods

This study was a pragmatic randomized controlled trial with pediatricians who registered for a four-day CPD conference – the SickKids Paediatric Update; an annual conference led by the Hospital for Sick Children and University of Toronto. Workshops and keynotes include a mix of didactic, interactive, and group based learning activities. Concurrent learning may include between 12 and 80 learners per workshop with typically over 300 paediatricians participating from across Canada.

Ethical approval for this study was provided by the SickKids Hospital Research Ethics Board. The study received funding from the Department of Pediatrics and the Education Development Fund at the University of Toronto.

Population

The participants in this study were paediatricians from across Canada. Upon on-line conference registration, electronically linked information about the study was provided and consent was sought with alternatives to participation given to all registrants (with options including full CME credit). Participants were able to choose particular workshops to attend during the conference.

Design

Upon electronic conference registration and workshop selection, consenting paediatricians, were randomly assigned based on a preset randomization sequence, to the intervention or control group. Paediatricians in the intervention groups received the MCQ based TEL interventions while the paediatricians randomized to control groups did not receive any testing or intervention. Since participants could choose multiple workshops, they were informed they may be in the intervention group (receiving TEL) for some workshops and control for the other. See Figure 1 for a summary of the design.

TEL Package

We endeavored to design the TEL package based on best principles described by Larsen (Larsen, et al., 2015) and on best practices in MCQ question writing. The faculty leads for each of the conference’s 16 workshops were asked to identify five ‘take-home’ key points of learning for their respective sessions. For each of the five key messages, we asked for two sets of corresponding multiple-choice questions (MCQs).

Only the 1st set was used as the TEL intervention, for both pre and post testing, while the 2nd set was used as an outcome measure. Our 2nd set of MCQs had a clinical vignette associated with each question, to allow us to extrapolate application of learning and to rule out the effect of simply learning the items and answers on the tests.

The TEL ‘intervention’ for each workshop consisted of two administrations of the same test. First, a pre-test of 5 MCQ questions was delivered 1 week before the CPD conference workshop. MCQs took the form of a single-best answer with 5 options per item and were framed as standalone items with a focus on the 5 key objectives for each workshop. Participants completing pre-test MCQs received no feedback.
However, results from the pre-tests were given to workshop leads to supplement learning needs assessments prior to leading workshops. Following the workshop, participants in the intervention group received the same 5 MCQs as post-tests two week later, but this time each item came with feedback. For each incorrect option, an explanation as to why the choice was incorrect was given automatically and participants were asked to make another choice. This step was repeated until the correct choice was made. We chose to limit the number items to five based on a previous feasibility pilot.

Neither pre nor post-tests were timed or modified in anyway. Each test was accompanied by a brief survey on the quality of questions, the impact on learning behaviour and an opportunity to provide feedback on the intervention package.

Faculty were invited to attend a best-practice MCQ development session based on MCQ writing principles (Case & Swanson, 1998) (Medical Council of Canada, 2010). Later in the year, the MCQs submitted were reviewed and, if necessary, edited by members of the research team experienced in test development to avoid cues that might render the MCQs less discriminating (e.g. grammatical cues, distractor length cues, logical cues etc.). MCQs were delivered electronically through the Qualtrics® survey platform.

Outcome

Our primary outcome was performance on the first attempt of the retention test for each workshop delivered 4 weeks after the conference. All participants received a common retention test for each workshop attended. The test consisted of 5 new MCQ questions in a vignette format (the second set of questions created by workshop leads). These questions were intended to evaluate retention and ability to use the knowledge of the major learning objectives of each workshop. We also calculated a performance score for each participant by calculating performance across all tests divided by the number of workshops attended.

Secondary outcomes included self-reported changes in behaviour of learners and teachers, perceptions of the utility of the TEL package including efficiency, effectiveness and satisfaction. This data was triangulated using objective data such as time taken and test completion. As this was a pragmatic trial, we could not control for amount of studying or review between sessions as well as other factors such as varying quality of workshop material and experience of presenters.

Data analysis

Since workshop participation size varied greatly and given that participants were randomly assigned to TEL or control across multiple workshops, we could not analyze our data using a traditional factorial ANOVA. Instead, we analyzed the data using a multi-level regression to account for the variable effect of the TEL intervention on participants and across workshops. Thus we analyzed with intervention as a fixed factor and participant (level 1) and workshop topic (level 2) as random factors. Additional covariates included post-test performance, inter-item reliability of the outcome test for the workshop, number of workshops attended and number of workshops attended in the intervention group. All analyses set the alpha threshold at 0.05 two-tailed. Where multiple analyses were performed, the Bonferroni correction was used to correct for the inflation of Type 1 error rate.

In a pragmatic trial, statistical significance is secondary to the size of the effect of the intervention (19). In order to summarize the effect size, we also analyzed the primary outcome for each workshop
measured as an individual ‘study’ and cumulatively compared performance for the intervention across workshops using a random-effects meta-analysis.

Secondary outcomes included performance on pre and post-test and responses to 3 satisfaction surveys at the end of the Pre, Post, and Retention Tests inquiring about participants perspectives on the utility of TEL. Results are reported as descriptive statistics or summaries of comments. All analyses were performed on SPSS 21 and SAS 9.2.

Results

Overall, 308 pediatricians registered for the conference and 186 individuals consented to participate across 15 workshops for which randomization to TEL or control occurred. For our primary outcome, complete data was available for 126 participants (68%) (445 individual data points). Workshop sizes varied from over 80 to 10 individuals per workshop with the final sample size for retention ranging from 50 to 13 per workshop.

The median number of different topic workshops that participants completed was 4 though the mode was 6. The range extended from 1 workshop to 8 with the inter-quartile range from 2 to 6 workshops attended by participants. Within our participant group 84% were randomized to receive the intervention for at least 1 workshop with 19% receiving it for two workshops, 27.3% for three workshops, 19% receiving the intervention for 4 or more workshops. The number of workshops attended and number of workshops attended randomized to the intervention were not significant correlates of participants’ aggregate performance on the retention tests (r=-0.15, p>0.05 and r=-0.14, p>0.05).

Average score across each workshop for the pre and post-tests is presented in Figure 2. Repeated measures ANOVA showed the aggregate increase from pre to post to be significant (F(2,174)=12.12, p<0.0001, partial eta squared = 0.14). Pre and post-performance was correlated 0.68 (p<0.0001) but post-test and retention were only weakly correlated at 0.21 (p<0.0001).

The primary analysis of the data showed a significant effect of TEL on retention performance for those in the intervention group compared to controls (Beta= 0.434, t=4.01, p<0.0001 (95% CI: 0.22-0.64)). Summarizing retention scores across participants and workshops, scores were significantly higher for TEL participants (71.2% correct choices) than those randomized to workshop-only (60% correct choices). Covariates including: number of workshops attended, number of workshops attended with intervention, number of workshops attended with control, and post-test performance were not significant. Retention performance showed significant clustering (Wald test Z=2.25, p<0.025) by individual performance (ICC=0.09) and by workshop topic (Wald Z=3.54, p<.003).

The random-effects meta-analysis allowed us to calculate a pooled-effect size indicating moderate effect of TEL (Hedge’s g of 0.46, 95% CI: 0.26-0.67) (see Figure 3) though the effect varied depending on the workshop topic as suggested by regression analysis.

The majority of participants (>80%) agreed that pre-tests helped identify knowledge gaps and enhanced learning at the workshop and moreover, that post-tests with feedback helped verify learning. Aggregate satisfaction data is reported in Figure 4.
Do one *then* see one: Sequencing discovery learning and direct instruction for simulation-based technical skills training

Introduction

When teaching technical skills, educators often include a mix of learner self-regulation and direct instruction. Appropriate sequencing of these activities - such as allowing learners a period of discovery learning prior to direct instruction - has been shown in other domains to improve transfer of learning. We compared the effect of allowing learners to try a novel simulated suturing task before formal instruction (Do then See) vs. the more typical sequence of formal instruction, followed by practice (See then Do) on skill retention and transfer.

Methods

First-year medical students (N=34) were randomized into two groups to learn horizontal mattress suturing. The See then Do group had access to instructors before independent practice, while the Do then See group explored the task independently before accessing instructors. Participants were assessed at the transition between interventions, and as training ended. One week later, we assessed their skill retention, and transfer to a novel variation of the suturing task. Performance was scored on a 5-point global-rating scale by a blinded rater.

Results

The groups did not differ significantly on immediate post-test or retention test \([F(1,30)=0.96, p<0.33]\). The Do then See group (N=16) outperformed the See then Do group (N=16) on the transfer test; 2.99 vs 2.52 \([F(1,28)=10.14, p<0.004, \eta^2 = 0.27]\).

Conclusion

Sequencing discovery and direct instruction optimally can improve transfer performance in simulation-based skills training. We discuss implications related to findings in other fields, future directions for research, and session and curricular design in the health professions.
Combining Scores based on Compensatory and Non-Compensatory Scoring Rules to Assess Resident Readiness for Unsupervised Practice: Implications from a National Primary Care Certification Examination in Japan

Purpose

Competence in health professions education requires combining scores from multiple sources and identifying pass-fail decisions based on compensatory and non-compensatory scoring decisions. However, guidance is lacking in the literature. This study uses national data to investigate consequences of combining scores, reliability, and implications for validity using a board certification examination with four subcomponent assessments.

Methods

We used national data from three consecutive years (2015, 2016, and 2017) of the Japan Primary Care Association Board Certification Examination, consisting of four subcomponent assessments—(1) Clinical Skills Assessment-Integrated Clinical Encounter (CSA-ICE), (2) CSA-Communication and Interpersonal Skills (CSA-CIS), (3) Multiple-Choice Questions (MCQ), and (4) Portfolio. Generalizability theory was used to estimate variance components and reliability. Kane’s composite reliability and kappa decision consistency were used to examine the impact of using compensatory and non-compensatory scoring approaches.

Results

Mean performance \((n=251)\) on the CSA-ICE, CSA-CIS, MCQ, and Portfolio subcomponent assessments were 61\% (SD=11\%), 67\% (SD=13\%), 74\% (SD=8\%), and 65\% (SD=9\%), respectively; component-specific \(\Phi\)-coefficient reliability ranged between .57-.67, .50-.60, .65-.76, and .87-.89, respectively. Using a completely non-compensatory scoring approach on all four subcomponents (i.e., pass each subcomponent individually), decision-consistency reliability was .33. Combining CSA components to yield three-component non-compensatory assessment resulted in reliability of .50. Fully-compensatory scoring yielded reliability of .86.

Conclusions

Assessing a wide range of abilities in making entrustment decisions requires considering the balance of assessment tools measuring distinct but related competencies. Educators should investigate the impact of non-compensatory scoring by examining its measurement characteristics, in addition to curricular, clinical, and patient safety considerations.