ALL-SAFE PRELIMINARY VALIDITY EVIDENCE
Laparoscopic Appendectomy Module 2

Analysis by DM Rooney (dmrooney@med.umich.edu) 10/7-30/22

Part A. Internal Evaluation of Measures of ALL-SAFE Laparoscopic Appendectomy (Lap Appy) Cognitive Testing Tool

METHODS

Study. 24 participants from 4 sites completed the web-based module. Participants included 15 novice, 6 intermediate, and 3 expert participants. All participating sites were represented (Mbingo, n=9; MRS, n=1; Soddo, n=6; UM, n=8).

Scoring and Statistical analyses.
The identical (but shuffled in presentation) 10-item pre- and post-module quizzes were scored dichotomously (1=correct, 0=incorrect) and summed for each participant, with a maximum score of 10. Pre- and post-module summed scores were compared using paired student-test, while differences between novice, intermediate, and expert participants was tested using one-way ANOVA, both with SPSS Statistics for Windows v.25 (IBM, Armonk, NY) Item-level analyses were performed using a many-facet Rasch model using Facets software v. 3.50 (Winsteps.com, Beaverton, OR) following anchoring on subjects to accommodate for nested design across sites.

RESULTS

Test of Score Change Following Training.
For all. Paired Student T-Test Comparison of pre- and post-intervention Quiz (Appendix X) summed scores from all 24 participants indicated that there was not a statistically significant improvement in mean summed scores from Pre (M=7.13, SD=1.6) to Post (M=7.15, SD = 1.9), p=.55
Rasch analysis at item-level was consistent with this finding, indicating no statistical difference from pre- (M=0.7) and post (M=.8) training, p=.99. Deeper analysis indicated score improvement for both novice and intermediate participants.

*****For novice participants.*** Paired Student T-Test Comparison of pre- and post-intervention Quiz summed scores from all 15 novices indicated that there was a statistically significant improvement in mean summed scores from Pre (M=5.87, SD=2.03) to Post (M=7.47, SD = 1.58), p<.001.

*****For intermediate participants.*** Paired Student T-Test Comparison of pre- and post-intervention Quiz summed scores from all 6 intermediate participants indicated that there was a statistically significant improvement in mean summed scores from Pre (M=7.83, SD=.98) to Post (M=9.50, SD = .84), p=.032.

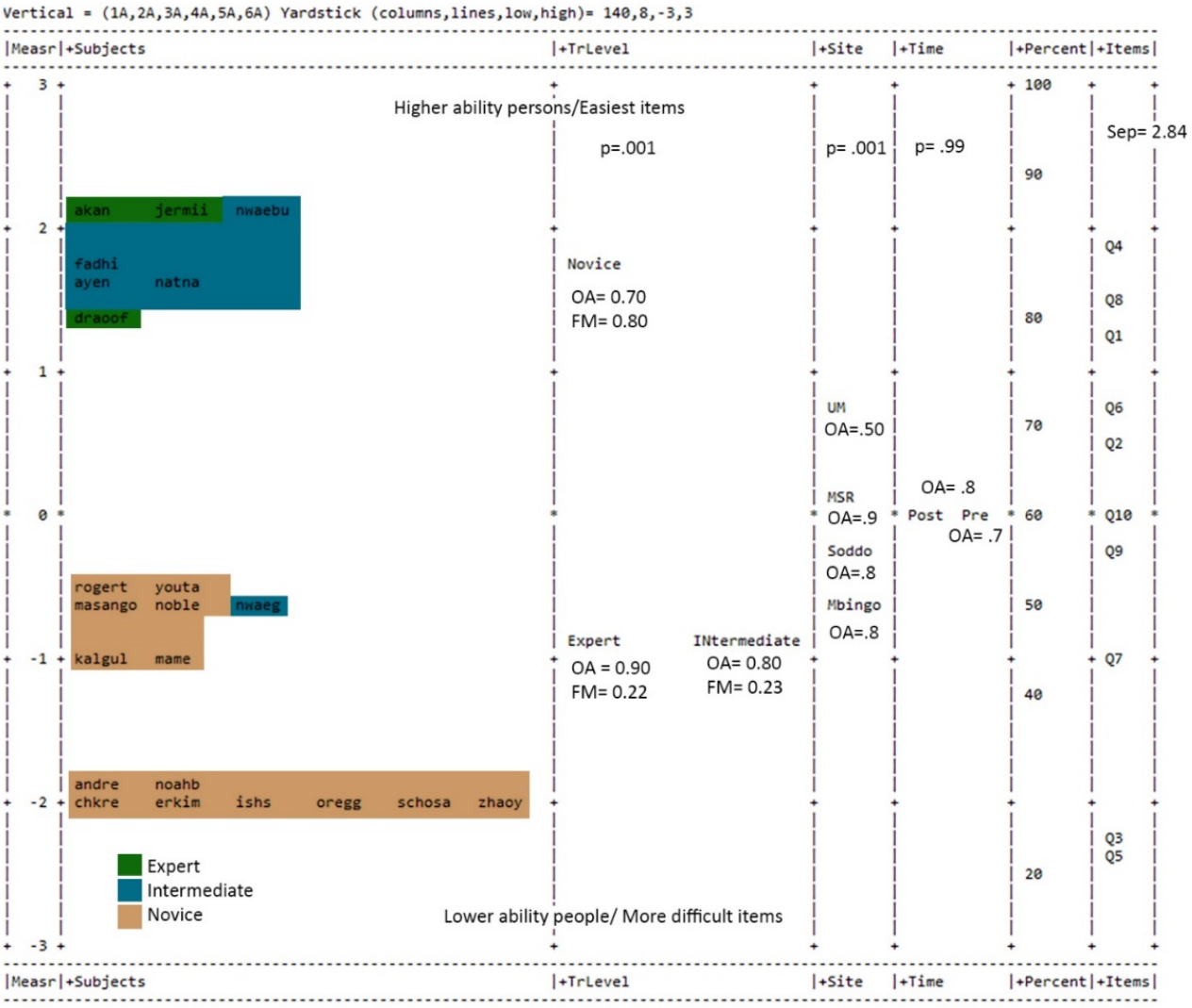
*****Discrimination across novice, intermediate, and expert participants.*** One-way ANOVA test indicated summed scores were able to discriminate between novice (M=6.67, SD=1.97), intermediate (M=8.67, SD=1.23), and expert (M=8.83, SD=.40) participants, p<.001. Rasch analysis supported this finding, X^2 (2,X)=70.8, p=.001

Bias analysis

One-way ANOVA indicated statistical differences in mean summed scores across sites, M_{UM}=5.38, M_{MSR}= 8.50, M_{Soddo} =8.08, M_{Mbingo}= 8.50), p <.001. Rasch analysis supported this finding, M_{UM}=0.5, M_{MSR}=0.9, M_{Soddo} =0.8, M_{Mbingo}=0.8), p=.001.

Although these findings could suggest item or test bias, because the majority of medical students (the lowest performers) in the UM cohort, differences can be rationalized.

Figure 1. Rasch Variable Map Cognitive Test, questions 1-10



Cognitive Test Item discrimination

Review of item discrimination showed reasonable distribution of item difficulty for items, with items 3 and 5 as the most difficult item (item discrimination=.15 and .67, respectively), and Qs 4 and 8 as the easiest (item discrimination= .81 and .86, respectively) (Table 1)

Table 1. Item discrimination values for ALL_SAFE cognitive test items, ordered highest to lowest.

Item No.	Item Difficulty	Estimated Discrimination*	Discrimination Power	Notes	Suggested Action
Q7	Moderate	1.66	High		—
Q6**	Easier	1.40	High	All Intermediate=1.0	—
Q1	Easiest	1.17	High	All Intermediate=1.0	—
Q2	Easier	1.16	High		—
Q9	Moderate	0.97	Good		—
Q10	Moderate	0.95	Good	Intermediate score declined after training (MPre=1.0, MPost=.83)	Review question for clarity/alignment with content. Focus groups with Residents to ID problem
Q8	Easiest	0.86	Good	All Intermediate=1.0	—
Q4	Easiest	0.81	Good	All Intermediate=1.0	—
Q5	Most Difficult	0.67	Low	All Intermediate=1.0 Novices; Remained difficult (Mpre=.27, Mpost=.53)	High Rasch MnSq Infit (1.67) suggests guessing from lower ability participants, so review question/content to ensure they align; review question to ensure clear
Q3	Most Difficult	0.15	Low	Novices; Remained difficult (Mpre=.13, Mpost=.47) Intermediate; No change in pre-post score.	Simply a difficult question that seems to be too hard for this targeted group of participants

*Values over 1 indicate this item has more discrimination power than expected for its difficulty while values under 1 indicate less discrimination power for its difficulty.

** Question 6 is the only question all novices answered correctly following training.

Considerations include:

a) Review/modification of Q5 to avoid ambiguity. Ensure question target is indeed covered within content.
b) Review Q3 to ensure it's clearly written and targeted content is covered
c) ** Q6 is the only question all novices answered correctly following training.
d) Likely, Intermediate participants came in with set knowledge (Pre-test means for Qs 1,4, 5, 6 = 1.0, SD=.00), which is expected
e) Given that mean post-test scores are still low (M=7.47, SD = 1.58) for novices, it might be expected that they review the content until they achieve mastery (100%) or some expected target, after ensuring content indeed aligns with QUESTIONS 3 and 5.

FINAL SUMMARY

Case Scenario/ Associated Cognitive Test:

- Cognitive test effectively discriminated between novice, intermediate, and expert participants, and demonstrated benefit to novice and intermediate participants with statistically significant score improvements for novice and intermediate groups, $p \leq .032$.
- Item discrimination analysis suggests review/potential modification of 2 questions (Qs 3/5) to ensure questions target is indeed covered within content, and language is clear.
- Evidence suggests Intermediate participants came in with set knowledge (Pre-test means for Qs 1,4, 5, and 6 = 1.0, SD=.00), which is expected
- Given that mean post-test scores are still low (M=7.47, SD = 1.58) for novices, it might be expected that they review the content until they achieve mastery (100%) or some expected target, after ensuring content indeed aligns with QUESTIONS 3 and 5.

Suggested Next Steps

1) evaluate 1) low, 2) borderline, and 3) high performer, with automated feedback based on AI to test alignment with scoring/competency decisions
2) Improve evaluation matrix to maximize distribution, minimize nesting which could introduce unexpected score patterns/biases.
3) To minimize future bias from experienced "novice" participants, recruit from new/virgin "novice" groups if possible, or include true experts as "gold standard"
4) Also, to avoid potential issues from nesting, consider recruitment of residents at UM/SUI, and if possible, novices (med students?) from all participating sites (Soddo/ Mbingo/ Kijabi) and ensure that each operator that submits a video is evaluated by a) judges from another site, b) these judges are ideally, balanced
5) To test judging quality of novice (medical students), add attendings to allow comparison to 'gold standards.'