ALL-SAFE PRELIMINARY VALIDITY EVIDENCE

Laparoscopic Appendectomy Module 2

Analysis by DM Rooney (dmrooney@med.umich.edu) 10/30/22-11/1/22

Part A. Comparison of Simulated Measures versus Live Measures

METHODS

Study. 2 Mbingo residents performed a live laparoscopic appendectomy within 2 weeks of each other at their home hospital. Their operative performances were videotaped. Eight evaluators (Mbingo, n=4, Soddo, n=4) representing 2 levels of experience (intermediate, n=7 and expert, n=1) reviewed the videos and rated performances using the 13-item VOP Checklist and 5-item Global Rating assessment tools. Comparison of simulated and live item-level, summed scores, and final rating was done using Kruskal-Wallis test in SPSS Statistics for Windows v.25 (IBM, Armonk, NY). Additional 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.

***Note: Inferences made from findings should be seen as preliminary trends, given reduced sample size
Reduced sample does not greatly impact overarching inferences but does impact some (Rasch) analyses. Notes indicate where this is the case.

RESULTS

Summary for Checklist

Although there were no statistical difference in performance ratings across the 2 trainees ($p>.05$), we highlight illustrate scoring patterns for both, treated separately.

Regarding Laparoscopic Appendectomy VOP Checklist and Global Ratings, Classical test analysis indicated (Table 1):

- There were no statistical score differences across simulated and live environments for the Laparoscopic Appendectomy VOP Checklist for any items for either trainee, p= |.06, 1.00|.
- There were no statistical score differences across simulated and live environments for the Laparoscopic Appendectomy VOP Checklist Summed Score, for either trainee, $p \geq .51$.
- There were no statistical score differences across simulated and live environments for the Laparoscopic Appendectomy VOP Global ratings for any items for either trainee, p= |.66, 1.00|.
- There were no statistical score differences across simulated and live environments for the Laparoscopic Appendectomy VOP Global Summed Score, for either trainee, $p \geq .72$.
- Final Rating (3-point: 1=Not competent, 2=Borderline, 3=Competent) were consistent for both trainees across simulated and live settings.

Table 1. Comparison of mean checklist performance ratings across simulated and live settings

item	Checklist item	TRAINEE 1 SIMULATED Mean (*SD*) n=8	TRAINEE 1 LIVE Mean (*SD*) n=8	P-value	TRAINEE 2 SIMULATED Mean (*SD*) n=8	TRAINEE 2 LIVE Mean (*SD*) n=8	P-value
1	Identifies anatomy of appendix, cecum and ileum by pointing *to each* with an instrument	2.00 (.00)	1.50 (1.00)↓	.69	2.00 (.00)	1.25 (1.04)↓	.06
2	Carefully grasps and elevates appendix	1.00 (1.16)	2.00 (.00)	.34	1.75 (.71)	1.75 (.71)	1.00
3	Mobilizes appendix by sharply taking down sidewall attachments	1.50 (1.00)	1.00 (1.16)	.69	2.00 (.00)	1.75 (.71)↓	.33
4	Avoids injury to appendix by excessive grasping or traction	2.25 (1.50)	2.25 (1.50)	1.00	1.88 (1.55)	2.63 (1.06)	.28
5	Creates window in mesoappendix bluntly by spreading with laparoscopic Maryland dissector	2.00 (.00)	1.00 (1.16)↓	.34	1.75 (.71)	1.50 (.92)↓	.55
6	Ligates appendiceal artery by placing figure of eight suture laparoscopically	.50 (1.00)	1.50 (1.00)	.34	1.75 (.71)	1.50 (.93)↓	.55
7	Performs intracorporeal knot with a surgeon's knot followed by two additional throws	2.00 (.00)	2.00 (.00)	1.00	2.00 (.00)	1.75 (.71)↓	.33
8	Avoids tearing the mesoappendix while placing ligating suture	3.00 (.00)	3.00 (.00)	1.00	2.25 (1.39)	2.44 (1.21)	.55
9	Cuts remainder of mesoappendix off of appendix using laparoscopic scissors	2.00 (.00)	1.00 (1.16)↓	.34	2.00 (.00)	1.75 (.71)↓	.33
10	Places two suture loops/endoloops at base of appendix	1.50 (1.00)	1.50 (1.00)	1.00	2.00 (.00)	1.75 (.71)↓	.72
11	Transects appendix sharply	2.00 (.00)	1.50 (1.00)	.69	2.00 (.00)	2.00 (.00)	1.00
12	Avoids leaving residual appendix on cecum (<3mm)	.75 (1.50)	2.25 (1.50)	.34	2.25 (1.39)	2.25 (1.39)	.33
13	Removes appendix from abdomen	2.00 (.00)	2.00 (.00)	1.00	2.00 (.00)	2.00 (.00)	1.00
–	**SUMMED**	22.50 (2.52)	22.50 (3.11)	.89	25.63 (2.67)	24.50 (3.85)	.51
item	**Global Domain**	TRAINEE 1 Simulated Mean (*SD*) n=8	TRAINEE 1 Live Mean (*SD*) n=8	P-value	TRAINEE 2 Simulated Mean (*SD*) n=8	TRAINEE 2 Live Mean (*SD*) n=8	P-value
1	Respect for Tissue	3.00 (.00)	3.25 (1.26)	1.00	3.63 (1.06)	3.38 (1.19)↓	1.00
2	Economy of Time and Motion	3.50 (1.29)	3.25 (.96)↓	.89	3.50 1.29)	3.08 (1.08)↓	.66
3	Instrument Handling	2.50 (.58)	3.00 (1.41)	.89	3.63 (1.30)	3.38 (1.06)↓	.68
4	Flow of Operation	3.75 (.96)	3.50 (1.29)↓	.89	3.50 (1.20)	3.75 (.89)	.64
5	Overall Performance	2.75 (.96)	3.00 (1.41)	.89	3.25 (.89)	3.38 (.92)	.79
–	**GLOBAL SUMMED**	15.50 (3.32)	16.00 (5.94)	.89	17.63 (4.98)	17.50 (4.63)	.96
–	TOTAL SUMMED	38.00 (3.56)	38.50 (5.75)	1.00	43.25 (7.25)	42.00 (6.41)	.72
	Final Rating	2.25 (.50)	2.25 (.50)	1.00	2.63 (.52)	2.63 (.52)	1.00

Supplemental Analyses for VOP Checklist and Global Ratings

Many-facet Rasch model which examined ratings differences using a 7-facet Rasch model (Trainee x Judge x Judge Expertise x Judge Site x Setting x Final Rating x Items).

VOP Checklist

Regarding Laparoscopic Appendectomy VOP Checklist, Rasch analysis was consistent with classical test analysis and indicated (Figure 1):

- There were no statistical rating differences across simulated versus live setting, p=.23
  Note: Model standard error = .34 [marginally close to reasonable thresholds (.30) to make these inferences]
  Additionally;
- There were no statistical score differences across judges, p=.08
- There were no statistical rating differences across judge experience (intermediate and expert judges), p=.28
- There were no statistical rating differences across judge sites, p=.28
- Given the provided sample, Competent (M=2.0) →Borderline (M=1.9), but statistical significance was not attained, p=.07
  Note: "Not Competent" ratings were never applied with this sampling.
- 10 Checklist items that are not "Avoid" continue to be low discriminators, likely simply because of scoring rubric (0=No/2=Yes) (See Figure 1)

VOP Global Ratings

Regarding Laparoscopic Appendectomy VOP Global, Rasch analysis was consistent with classical test analysis and indicated (Figure 1):

There were no statistical Global rating differences across simulated versus live setting, p=.97, suggesting Global ratings could be used across settings. Note: Model standard error = .22 [within reasonable thresholds (.30) to make these inferences]
Additionally;

- There were statistical score differences across judges, p=.01, However, deeper bias analysis were inconclusive because the associated standard errors were too high (>.30), ranging .37-1.46, so indefensible (likely due to small sample size)
- There were no statistical rating differences across judge experience (intermediate and expert judges), p=.48, suggesting variability due to judge experience did not introduce rating bias.
- There were no statistical rating differences across judge sites, p=.13, indicating there was no rater bias across site.
- Given the provided sample, Competent (M=2.0) →Borderline (M=1.9), but statistical significance was attained, p=.01, suggesting the Final rating could successfully discriminate competent versus borderline performances. Care should be taken with this particular finding as variability in this sample of ratings could be from a number of unaccounted sources.
  Note: "Not Competent" ratings were never applied with this sampling.

Figure 1. Variable Map, Rasch Analyses Laparoscopic Appendectomy VOP Checklist

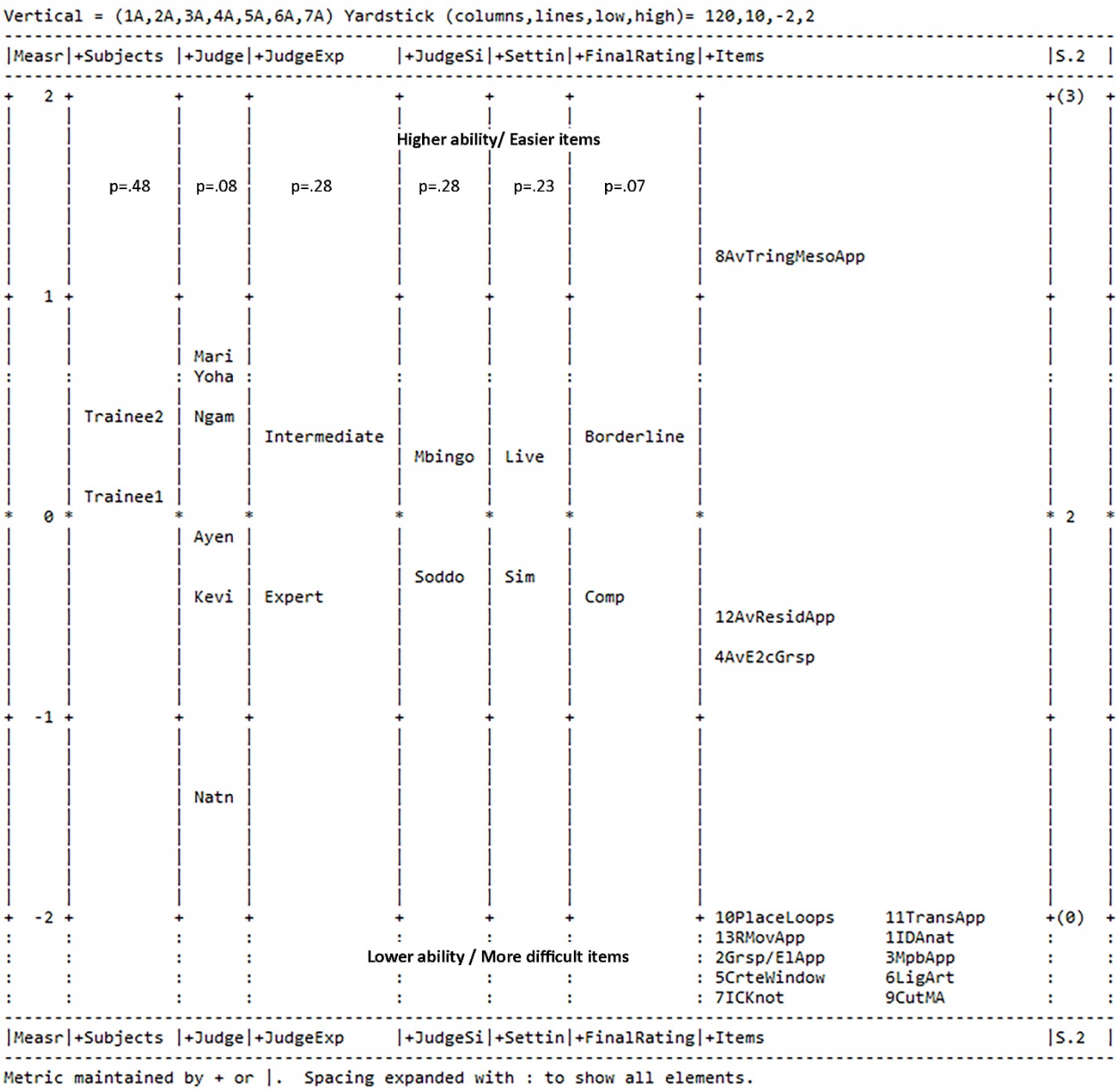
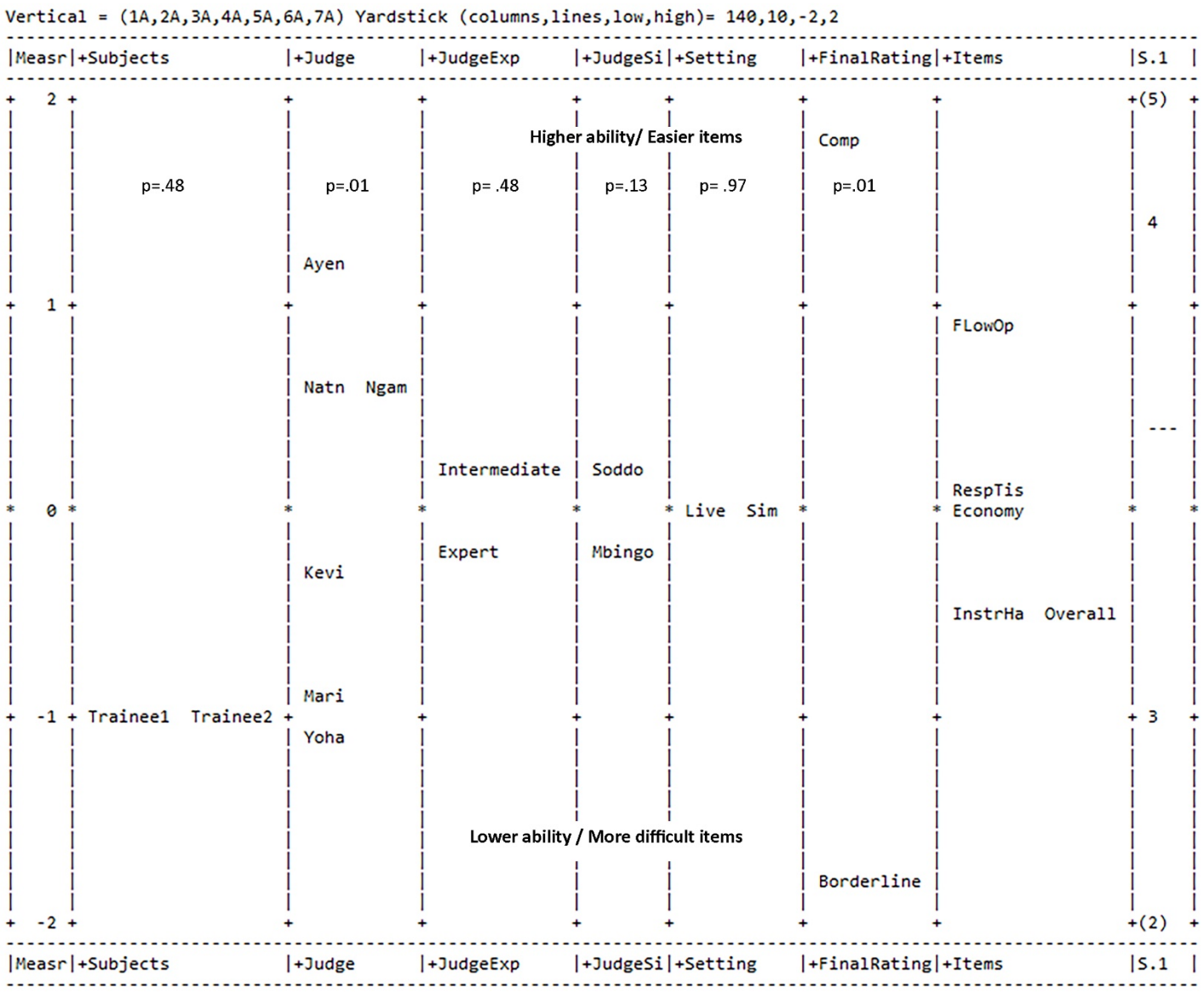


Figure 2. Variable Map, Rasch analyses Laparoscopic Appendectomy VOP Global Ratings



SUMMARY

Psychomotor Performance Checklist VOP Assessment

- Met intention to test Checklist function across simulated and live settings- indicated the Laparoscopic Appendectomy VOP Checklist was consistently used across settings.
- Although it has been previously established that the Checklist is less informative for performance assessment than Global ratings, we should still consider any outstanding issues, such as odd rating patterns we could explain.
- Looking at response patterns, both Trainees had 7/13 items (53%) with decreased scores when transitioned to the live setting.
- Of these items,
  Q1 (Identifies anatomy of appendix, cecum and ileum by pointing *to each* with an instrument),
  Q3 (Mobilizes appendix by sharply taking down sidewall attachments),
  Q5 (Creates window in mesoappendix bluntly by spreading with laparoscopic Maryland dissector), and
  Q9 (Cuts remainder of mesoappendix off of appendix using laparoscopic scissors) were lower in the live setting for both trainees.
  These decreases for this trainee can be explained primarily by the difficulty of this particular operative case (rated as a 5-'Most difficult case' by the overseeing attending), which was substantially more difficult that the "average" anatomy that was experienced in training. Additionally, differences in instrumentation, and need/ability to apply tasks to operative setting may have influenced *live* ratings.

Psychomotor Performance Global VOP Assessment

- Met intention to test Global function across simulated and live settings indicated the Laparoscopic Appendectomy Global ratings were consistently used across settings.
- Looking at response patterns, Trainee 1 improved from simulated to live Global ratings. Trainee 2 had 3/5 items (60%) with decreased scores when transitioned to the live setting.
  As explained previously, these differences in global ratings might be explained by the technical difficulty of the case presented in operative setting.

Suggested Next Steps

1) Check correlation of these Checklist summed scores and Global ratings to those captured via Artificial intelligence (AI).
2) Ongoing re-evaluation after collecting more data, keeping consideration of points 2-4 below.
3) Improve evaluation matrix to maximize distribution, minimize nesting which could introduce unexpected score patterns/biases.
4) To minimize future bias from experienced "novice" participants, recruit from new/virgin "novice" groups if possible, or include true experts as "gold standard"

5) 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
6) To test judging quality of novice (medical students), add sampling of attendings (from 3 sites) to allow comparison to 'gold standards.'