ALL-SAFE PRELIMINARY VALIDITY EVIDENCE

Analysis by DM Rooney (dmrooney@med.umich.edu) 8.9-8.12.21
Following removal of 2 UM judges familiar with trainees to minimize potential rater bias
Validity Evidence for use of ALL-SAFE VOP with Checklist (11 items) and Global (5 items)

Internal Structure: Comparison of Novice v. Expert Performance Ratings

Description: Comparison of performance scores across 14 performances (Novice, n=9; Expert, n=5) performances using Kruskal-Wallis (non-normal distribution confirmed) x 12 judges. Done in two ways; 1) *all raters* (n=12; including novices) and 2) *only expert*s (n=5; only attending) raters. *Scoring:* Checklist items 1,2,4,5,8,9,10, 11 rescored max score=2. Items 3,6,7 max score=3. Global items all max score=5. Max comb sum=50.

item	Checklist item	Novice Mean (*SD*) Combined All Raters	Experts Mean (*SD*) All Raters	P-value	Novice Mean (*SD*) Combined Expert Raters Only	Experts Mean (*SD*) All Raters Expert Raters Only	P-value
1	Evaluates both fallopian tubes by pointing to both with an instrument	1.67 (.75)	1.90 (.44)	.055	1.57 (.84)	1.91 (.42)	.094
2	Stabilizes involved fallopian tube by grasping adjacent to ectopic pregnancy site	1.82 (.58)	1.93 (.36)	.225	1.91 (.42)	1.91 (42)	.975
3	Avoids excessive grasping of fallopian tubes	2.68 (.94)	2.80 (.75)	.473	2.74 (.86)	3.00 (.00)	1.62
4	Creates a longitudinal salpingostomy	2.00 (0.0)	2.00 (.0)	1.0	2.00 (.00)	2.00 (.00)	1.0
5	Extends salpingostomy to encompass length of ectopic pregnancy	1.61 (.80)	1.90 (.44)	.019	1.39 (.94)	1.91 (.42)	.025
6	Avoids transecting involved fallopian tube	3.0 (0.0)	3.0 (.00)	1.0	3.00 (.00)	3.00 (.00)	1.00
7	Avoids damaging mesosalpinx when performing the salpingostomy	2.79 (.78)	2.90 (.54)	.379	2.87 (.63)	3.00 (.00)	.328
8	Evacuates at least 80% of ectopic contents from tube	1.11 (1.00)	1.43 (.91)	.091	1.04 (1.02)	1.48 (.90)	.160
9	Retrieves specimen from abdomen with laparoscopic instrument	1.71 (.71)	1.93 (.36)	.042	1.65 (.78)	2.00 (.00)	.043
10	Places single suture at marked edge of fallopian tube	1.75 (.67)	1.73 (.69)	.893	1.83 (.58)	1.83 (.58)	.963
11	Performs intracorporeal knot with a surgeon's knot followed by two additional throws	1.68 (.74)	1.97 (.26)	.007	1.65 (.78)	2.0 (.00)	.043
–	SUMMED	21.79 (2.81)	23.50 (1.88)	.001	21.65 (2.84)	24.04 (1.33)	.002

item	Global Domain	Novice Mean (*SD*) Combined All Raters	Experts Mean (*SD*) All Raters	P-value	Novice Mean (*SD*) Combined Expert Raters Only	Experts Mean (*SD*) Expert Raters Only	P-value
1	Respect for Tissue	3.22 (1.01)	3.89 (.96)	<.01	3.39 (1.16)	4.09 (.79)	.039
2	Economy of Time and Motion	2.78 (1.18)	3.63 (1.03)	<.01	2.65 (1.27)	3.48 (1.04)	.041
3	Instrument Handling	2.79 (1.21)	3.72 (1.06)	<.01	2.65 (1.27)	3.74 (1.01)	.004
4	Flow of Operation	3.37 (1.04)	4.14 (.92)	<.01	3.26 (1.14)	4.09 (.95)	.016
5	Overall Performance	2.59 (1.01)	3.42 (1.04)	<.01	2.61 (1.20)	3.57 (.99)	.006
–	GLOBAL SUMMED	14.75 (4.83)	18.79 (4.51)	<.01	14.57 (5.27)	18.96 (4.16)	.006
–	TOTAL SUMMED	36.79 (6.67)	42.13 (5.37)	<.01	36.22 (7.47)	43.00 (4.46)	.001
	Final Rating	2.30 (.78)	2.75 (.51)	<.01	2.22 (.85)	2.83 (.49)	.006

Summary. Findings suggest that comparison of novice /expert performance ratings at the item-level was not helpful to discriminate performance levels. In spite of this, summed total of the checklist did discriminate novice and expert performances, regardless of judge expertise. Further, global ratings at the domain-level were able to discriminate novice versus expert performances, as were the global summed and total (checklist and global summed. This suggests that the summed checklist and global ratings could be used to discriminate novice and expert performances.

Supplemental Analyses:

Many-facet Rasch model which examined ratings differences using a 6-facet Rasch model (ID x Subject Expertise x Institution X Judge Expertise x Final Rating x Item) indicated significant ratings differences across *Subject Expertise* facet, with Novice subjects' performance ratings (M=3.00) statistically lower that Expert subjects' performance ratings (M=3.7), X^2 (2,114)= 96.9, p=.001, suggesting ratings were able to discriminate between Novice and Expert performances.

The same Many-facet Rasch Model was used to examine ratings differences across *Final Rating* (e.g. Competent, Borderline, and Not Competent) response options. This analysis indicated statistical ratings differences across *Final Ratings*, shown below;

> Competent (M=4.0) →Borderline (M=2.5) → Not Competent (M=2.1)
> X^2 (2,114)= 455.5, p=.001, suggesting that these three response options could adequately discriminate subjects.

Internal Structure: Rater agreement across novice and expert judges

Description: Review of inter-rater agreement of 10 performances* across Novice (n=7) and Expert (n=5) raters, measured by averaged two-way mixed Intraclass correlation
(*selected for completeness of data, represented at least 3 from each site)

item	Domain	ICC	95% Confidence Interval
Checklist			
–	**Checklist Summed**	.96	.85 - .95
Global			
1	Respect for Tissue	.90	.70 - .90
2	Economy of Time and Motion	.90	.83 - .95
3	Instrument Handling	.90	.82 - .94
4	Flow of Operation	.89	.80 - .94
5	Overall Performance	.77	.58 - .87
–	**GLOBAL SUMMED**	.93	.88 - .96
–	**TOTAL SUMMED**	.95	.91 - .97
–	**Final Rating**	.88	.79 - .94

Summary: There were a good amount of rater agreement across novice and expert judges The lowest ICC value was estimated to be .77, for *Global-Overall Performance* item, suggesting moderate agreement between novice and expert judges for that item. Remaining ICC values ranged between .88 and .96, suggesting excellent interrater agreement for those items.[1]

Supplemental Analyses: Many-facet Rasch model which examined ratings differences using a 6-facet Rasch model (ID x Subject Expertise x Institution X Judge Expertise x Final Rating x Item) indicated no statistical ratings differences across *Judge Expertise* facet, with Trainee ratings (M=3.4) not statistically different than Expert ratings (M=3.4), X^2 (2,114) = 1.9, p = .16.

1. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

Internal Structure: Comparison of performance ratings across 3 sites

Description: Comparison of ranked ratings of same 10 performances across 3 sites (UMichigan, Mbingo, and Soddo) using Kruskal-Wallis test to ensure generalizability of scoring across sites

item	Domain	Mbingo (n=2) Mdn, |Range|	Soddo (n=7) Mdn, |Range|	UMichigan (n=5) Mdn, |Range|	P-Value	
Checklist						
–	**Checklist Summed**	23.00, |19,25|	23.00, |16,25|	24.00, |14,25|	.375	
Global					–	
1	Respect for Tissue	4.0, |2,5|	3.0, |2,5|	4.0, |1,5|	.128	
2	Economy of Time and Motion	3.5, |1,5|	3.0, |1,5|	3.5, |1,5|	.917	
3	Instrument Handling	4.0, |1.5|	3.0, |2,5|	3.0, |1,5|	.675	
4	Flow of Operation	5.0, |4.5|	4.0, |2,5|	3.5, |1,5|	**.002**	
5	Overall Performance	3.5, |2,5|	3.0, |1,5|	3.0, |1,5|	.162	
–	**GLOBAL SUMMED**	20.0, |10,25|	16.0, |7,25|	17, |5,25|	.268	
–	**TOTAL SUMMED**	43.0, |29,50|	39.0, |24,50|	39.0, |20,50|	.199	
–	**Final Rating**	3.0, |1,3|	3.0, |1,3|	3.0, |1,3|	.568	

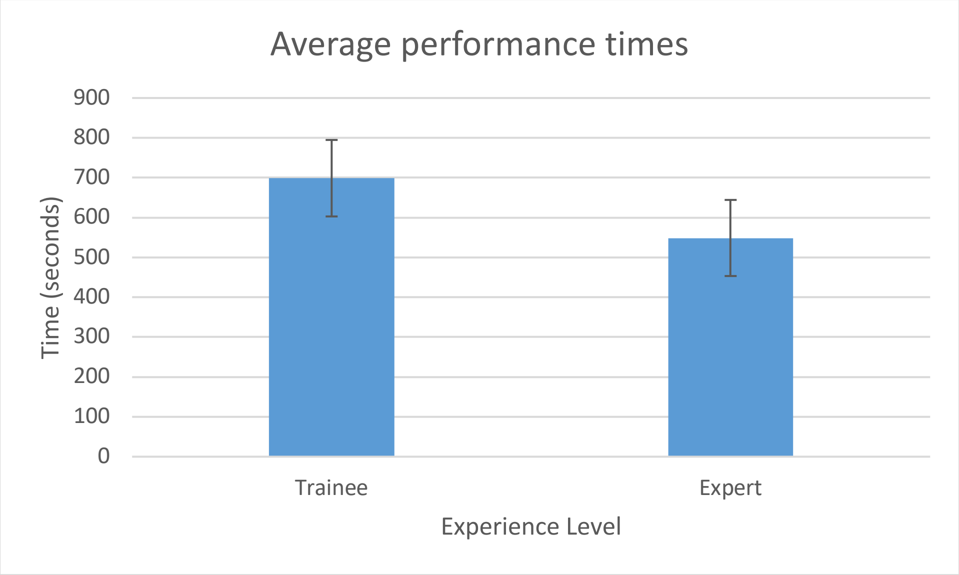
Summary: This supported preliminary evidence of generalizability of scoring across participating sites.

A Kruskal-Wallis test was conducted to evaluate scoring differences among the three sites (Mbingo, Soddo, and University of Michigan) for the Summed Checklist, each of the 5 Global items (domains), the Global summer score, total summed score (combines the checklist summed score and the Global summed score), and finally, the final overall rating. For all but one item (Global rating #4 Flow of operation), the test, which was corrected for tied ranks, was not significant, p = | .128,.568|, suggesting the proportion of variability in the ranked dependent variable (score) was not accounted for by participating sites, indicating a little relationship between site and the test scores. This supported preliminary evidence of generalizability of scoring across participating sites.

Supplemental Analyses: Many-facet Rasch model which examined ratings differences using a 6-facet Rasch model (ID x Subject Expertise x Institution X Judge Expertise x Final Rating x Item) indicated statistical ratings differences across *Institutions* with Mbingo (M=3.6) having statistically higher ratings than U Michigan (M=3.2) and Soddo (M=3.3), X^2 (2,114) = 74.8, p=0.001, suggesting rater training will be important to ensure rating calibration across institutions.

Internal Structure: Comparison of Novice v. Expert Performance Times

Description: Comparison of performance times across Novice (n=5) and Expert (n=5) performances using independent t-test (normality confirmed using both Kolmogorov-Smirnov and Shapiro-Wilk tests, p≥ 0.20).



Summary. Although average Trainee times were higher than Expert times, findings suggest that comparison of novice /expert performance times for this cohort were not helpful at discriminating performance levels.
M_{Novice} = 698.80 seconds (*SD*=214.6)
M_{Expert} = 548.80 seconds (*SD*=213.5)
P = .30 (NOT STATISTICALLY SIGNIFICANT)

This exercise pointed out an issue with the checklist, along with a few considerations re: timing. These are listed below, with potential solutions.

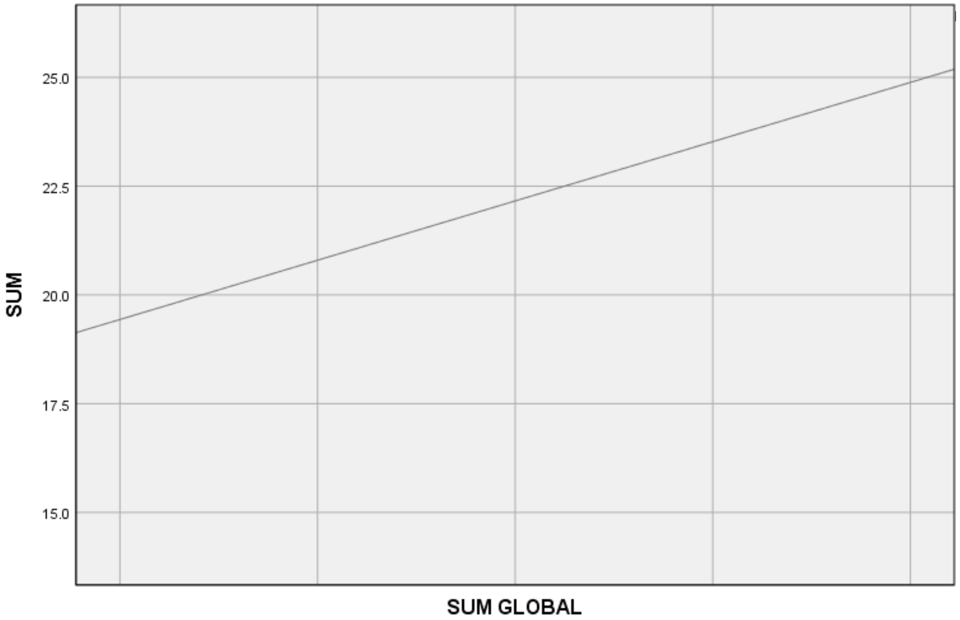
1. Many examinees did not cut suture at end of task, making it difficult to tell when task was done.
   Solution: Add "Cut suture tail on competing knot" to checklist
   Justification: Allows hard time stamp to be considered (from first touch of fallopian tube to cutting of suture)
   *For the sake of this exercise, I timed all from first touch to cinching of knot for consistency.
   Many examinees did not cut suture at end of task, making it difficult to tell when task was done.
2. Many examinees left substantial amount of ectopic contents.
   Could not adequately visualize to estimate amount (%) left, compromising associated item on checklist, and made performance times shorter for those examinees that chose not fully empty contents (biased benefit)

Solution: Should consider adding time penalties for this (do not have good example form this sample, so can't estimate time penalty) justification: FLS standard does this.

3. <u>Some examinees did not examine fallopian tube.</u>
   Not having a clear "start time" makes it difficult to measure, and erroneously shortens time for those that did not perform this task (biased benefit)
   Solution: Add time penalty (average time was 5-7 seconds in sample)

Relationships to Other Variables: Correlation of summed ALL-SAFE checklist scores with OSATS summed scores

Description: Correlation of summed checklist scores with OSATS scores (n=10) estimated by Pearson's r.



Summary. Findings suggest a positive correlation between summed ALL-SAFE Checklist score (SUM) and established OSATS summed score (SUM GLOBAL), r(114) = .534, p=.0001, supporting use of ALL-SAFE checklist summed score to measure performance skill.

Similarly, these summed scores (SUM) correlated with the Combined summed score of the Checklist and OSATS, r(114) = .96, p=.0001), as well as the Final Rating, scored on 3-point scale (1 = Does not demonstrate competence, 2 = Borderline, 3 = Does demonstrate competence), r(114)=.76, p = .0001.

Validity Evidence Relevant to Test Content

Full description of development process of associated curriculum materials and assessment tools to ensure transparency, and initial validation processes used with summary of findings

Case scenario and associated questions:

1. Originally drafted by a PGY2 resident (Freneh, Soddo Ethiopia) to ensure relevance to targeted learners
2. Reviewed by 2 Co-Is (Jeffcoach, GenSx; Marzano, OBGyn; UM) for content and relevance.
3. Reviewed by 2 M4 students-research assistants (Mott/ Yoonhee, UM) for clarity, content and flow,
4. Reviewed by a native English speaker copy-editor for clarity and grammar
5. Reviewed by psychometrician and (Rooney, Obayemi, UM) for stem clarity, response option bias, for content clarity, relevance, alignment of questions with scenario, and flow, based on 7/17 Self-assessment workshop (O'Keeffe; Royal College of Surgeons in Ireland)
6. Disseminated to entire research team for review
7. Final review and approval by PI/Co-I (Kim, UM; Barnard, SIU)

Assessment Tool: Pre/post Multiple Choice Quiz:

1. Originally drafted by PGY1 resident (Hsu, UMich), based on the case scenario.
2. Reviewed and edited for relevance/content by PI (Kim)
3. Reviewed by 2 M4 students-research assistants (Mott/ Yoonhee) for clarity.
4. Reviewed by psychometrician and PGY2 resident (Rooney, Obayemi) for stem clarity, response option bias, for content clarity, relevance, and flow.
5. Disseminated to entire research team for review
6. Final review and approval by PI (Kim)

Continued on next page

Assessment Tool: Performance Assessment (Verification of Proficiency):

1. Drafted by CO-I (Barnard)
2. Reviewed by entire research team for content and relevance.
3. Dissemination to sites for trial of practical use with performance videos (change language form salpingectomy to salpinostomy
4. Disseminated to entire research team for review
5. Review with Global Surgical Training Challenge assessment expert (Dara O'Keeffe, Royal College of Surgeons in Ireland), 6/17/21
5. Review with ALL-SAFE research team to discuss proposed changes based on 6/17 suggestions
6. Edited by PI/Co-I to split 1 item (item 3), and add 3 additional "error-based items," and split of final designation to "Competent, Borderline, Not competent "
7. Review by psychometrician (Rooney) for clarity, relevance, alignment of questions with skills
8. Captured data with 10 performances (2N/2E from 3 sites (Soddo, Mbingo, UM) x judges/raters, indicated need for additional item "Cut suture" to allow an actual end time to be observable.

Materials: ALL-SAFE Box Trainer

1. Development of ALL-SAFE box trainer was done by team of 3 Engineer students (3DI Lab, UMich) following parameters of PI (material = cardboard, functions and approximate dimensions of avg lap. pelvis/abdomen field)
2. Initial prototype trialed by PI/CO-Is (Kim, Barnard, Jeffcoach, Snell) for ease of build, stability, lighting.
3. Expanded trial: 9 laparoscopic surgeons (2 Gen Surgery residents, 4 General Surgery attendings, 3 Ob/Gyne attendings) trialed the box. Four participants evaluated the box trainer's ease of build, while all 9 evaluated the box trainer's characteristics, and the ectopic pregnancy simulator, using 2 checklists;
    a. <u>Ease of build (5-item 5-point rating scales)</u>
       - Ease of Build M ratings range = |4.33, 4.50|, with no suggestions for improvements
    b. <u>Box Trainer Attributes (6-item 5-point rating scales)</u>
       - Box Trainer Attributes ratings range = |2.50, 3.25|, with suggestions for improvements including "Need to readjust the port sites", which targeted making the port sites bigger to better align with real surgical experience

       ACTIONS TAKEN:

       1) Increasing the internal scaling to improve the view
       2) Increase scale, and placement of the "port holes" to better align with authentic surgical experience

Materials: ALL-SAFE Simulated Uterus/Ectopic Pregnancy

1. Development of simulated uterus/ectopic pregnancy was done CO-I (Barnard)
   (parameters = representative of relevant anatomy, low-cost, accessible materials)
2. Initial prototype trialed by PI/CO-Is (Kim, Barnard, Jeffcoach, Snell)
3. Expanded trial: 9 participants (2 Gen Surgery residents, 4 General Surgery attendings, 3 Ob/Gyne attendings) trialed the simulated ectopic pregnancy. All participants evaluated the associated build directions (ease of build) and the simulator's characteristics, using 3 checklists;
   a. <u>Ease of build (5-item 5-point rating scales)</u>
      - Ease of Build M ratings range = |4.00, 4.50|, with no suggestions for improvements
   b. <u>Ability to Perform Task (5-item 5-point rating scales)</u>
      -Ability M rating range = |2.00, 3.86|
   c. <u>Simulator Attributes (13-item 5-point rating scales)</u>
      - Box Trainer Attributes ratings range = |1.00, 5.00|, with suggestions for improvements targeted 2 primary areas;
      a) Ectopic pregnancy, with comments that included:
         "Need to have something a bit more solid for the ectopic,"
         "Would suggest thicker substance such as playdoh to mimic clot/tissue of ectopic"

      b) Fallopian tube, with comments that included:
         "Typically fallopian tube has more resistance than penrose, so penrose was easier to cut"

      ACTIONS TAKEN:

      1) Change ectopic pregnancy contents to play dough recipe
      2) No changes to fallopian tubes as no viable alternative to penrose drain that is low cost was identified