INTERNAL EVALUTION OF ALL-SAFE PROGRAM VOP PERFORMANCE MEASURES

Analysis by DM Rooney (dmrooney@med.umich.edu) 7.2.23 - 7.6.23

Validity Evidence for use of 4 ALL-SAFE VOPs with Checklist (variable items items) and Global (5 items) Combined as a single program
Construct Measured: Laparoscopic Surgical Skills

Description: Comparison of all performance scores across trainee groups (Medical students, n=10; Resident, n=23; Faculty, n=5) performances using Kruskal-Wallis (non-normal distribution confirmed). Caveat- not all operators, and not all raters completed each module so these are not fully linked. Because of this, independent samples analysis was performed. *Scoring:* Because all checklists had different number of items, Checklist Summed and Total Summed scores were normalized (out of 100%). Global items all max score=5, max Global sum=25

Internal Structure: Comparison of Novice v. Intermediate v Experienced Performance Ratings

Table 1. Comparison of performance ratings across participant groups, medical student, resident, and faculty

item	Global Domain	Medical Students (n=59) Mean (*SD*) All Raters	Residents (n=225) Mean (*SD*) All Raters	Faculty (n=56) Mean (*SD*) All Raters	P-value
	Checklist SUMMED*	65.14 (21.53)	72.31 (16.28)**	76.51** (15.35)	.008
1	Respect for Tissue	2.68 (.97)	3.28 (.85)	4.32 (.72)	<.001
2	Economy of Time and Motion	2.14 (1.04)	3.00 (.93)	4.13 (.81)	<.001
3	Instrument Handling	2.10 (1.06)	3.07 (.91)	4.25 (.88)	<.001
4	Flow of Operation	2.58 (1.05)	3.40 (.89)	4.43 (.81)	<.001
5	Overall Performance	2.03 (.87)	3.03 (.86)	4.00 (.85)	<.001
–	GLOBAL SUMMED	11.53 (4.39)	15.78 (3.91)	21.13 (3.49)	<.001
	Normalized Global Summed	**46.10 (17.55)**	**63.13 (15.66)**	**84.50 (13.96)**	<.001
–	**TOTAL SUMMED** *	**56.20 (16.36)**	**70.04 (13.30)**	**78.47 (11.22)**	<.001
	Final Rating (max=3.0)	2.49 (.60)	2.35 (.74)	2.34 (.72)	.50

* (normalized out of 100%)
** No statistical difference between medical students/faculty

Summary
Findings suggest that comparison of novice (medical student) /intermediate (resident)/ and faculty (experienced) performance ratings for all measures, exception Final rating, discriminated Novice v. Intermediate v. Experienced performances.

Internal Structure: Comparison of Performance Ratings over Time

Description: Comparison of all performance scores across time (order of data collection (Module 1=Time 1, Module 2=Time 2, and Modules 3 and 4 = Time 3) for Medical students' (n=10) and Residents' (n=23) performances using Kruskal-Wallis (non-normal distribution confirmed via SPSS v28.0). Caveat- not all trainees and raters completed each module so these are not fully linked. Because of this, independent samples analysis was performed. *Scoring:* Because all checklists had different number of items, Checklist Summed and Total Summed scores were normalized (total percent out of 100%). Global items (items 1-5) all max score=5, for a maximum Global Summed =25

Table 2. Comparison of performance ratings over time

item	Global Domain	TIME 1 (n=95) Mean (*SD*) Trainees Only	TIME 2 (n=81) Mean (*SD*) Trainees Only	TIME 3 (n=108) Mean (*SD*) Trainees Only	P-value
	Checklist Summed*	60.88 (7.52)	78.29 (19.92)^	82.71 (16.51)^	<.001
1	Respect for Tissue	3.24 (.96)^	2.84 (.87)	3.31 (.84)^	<.001
2	Economy of Time and Motion	2.86 (1.11)^	2.49 (.99)	3.04 (.90)^	<.001
3	Instrument Handling	2.89 (1.12)	2.58 (1.00)^^	3.06 (.90)^^	.05

4	Flow of Operation	3.49 (1.01)^	2.77 (.97)	3.33 (.84)^	.001
5	Overall Performance	2.65 (.95)^	2.59 (.97)^	3.15 (.85)	<.001
–	GLOBAL Summed	15.15 (4.53)^	13.27 (4.32)	15.90 (3.92)^	.002
	Normalized Global Summed	**60.59 (11.27)^**	**53.09 (17.30)**	**63.59 (15.67)^**	<.001
–	**TOTAL SUMMED** *	**60.74 (11.27)^^**	**66.62 (16.56)^^**	**73.23 (14.44)^^**	<.001
	Final Rating (max=3.0)	2.49 (.67)^	2.41 (.69)^	2.25 (.75)^	.054

* (normalized out of 100%)

^ No statistical difference between time points

^^ Statistical difference between time points

Table 3. Comparison of performance ratings across modules.

Summary

All scores were able to discriminate learners' scores over time, exception Final Rating.

item	Global Domain	Module 1 (n=6, n=4) Mean (*SD*) Trainees | Faculty	Module 2 (n=17, n=1) Mean (*SD*) Trainees | Faculty	Module 3 (n=12, n=1) Mean (*SD*) Trainees | Faculty	Module 4 (n=15, n=1) Mean (*SD*) Trainees | Faculty	P-Value	η2
–	Checklist SUMMED*	60.49 (7.49) | 63.41 (6.09)	78.29 (19.92)|97.41 (5.17)	81.01 (18.55)| 94.23 (11.54)	84.07 (14.70)| 97.00 (6.00)	<.001	.42
1	Respect for Tissue	3.23 (.97) |4.09 (.92)	2.84 (.87)|4.25 (.50)	3.19 (.87)|4.25 (.96)	3.42 (.81)|4.50 (.58)	<.001	.12
2	Economy of Time and Motion	2.81 (1.15)|3.85 (.92)	2.49 (.99)| 4.75 (.50)	2.96 (.90)|4.25 (.96)	3.10 (.90)|4.25 (.50)	<.001	.10
3	Instrument Handling	2.83 (1.16)| 3.98 (96)	2.58 (1.00)| 5.0 (.0)	3.00 (.90)|4.25 (.96)	3.12 (.90)|4.25 (.50)	.04	.08
4	Flow of Operation	3.44 (1.01)|4.28 (.90)	2.77 (.97)|5.0 (.0)	3.31 (.83)| 4.25 (.96)	3.35 (.86)| 4.50 (1.00)	<.001	.17
5	Overall Performance	2.63 (.98) | 3.63 (1.01)	2.59 (.97) |4.25 (.50)	3.04 (.87)|4.25 (.96)	3.23 (.83)|4.50 (.58)	.04	.08
–	GLOBAL SUMMED	14.95 (4.67) | 19.83 (4.17)	13.27 (4.32)|23.45 (.5)	15.50 (3.96)|21.25 (4.50)	16.22 (3.89)|22.00 (2.00)	<.001	.13
–	**Normalized Global Summed**	59.80 (18.69)|79.33 (16.66)	53.09 (17.30)| 93.0 (2.0)	62.00 (15.83)|85.00 (18.00)	64.87 (15.55)| 88.80 (8.00)	<.001	.13
–	**TOTAL SUMMED** *	60.15 (11.61)|(71.37 (9.43)	66.62 (16.56)|95.37 (3.21)	71.69 (15.62)| 89.71 (14.44)	74.47 (13.43)| 92.50 (4.12)	<.001	.14
–	Final Rating (max=3.0)	2.49 (.65)| 2.33 (.75)	2.41 (.69)|3.0 (.0)	2.25 (.75)|3.00 (.00)	2.35 (.71)|2.00 (.82)	.18	—

* (normalized out of 100%)

Internal Structure: Comparison of Performance Ratings and Normalized Scores over Module

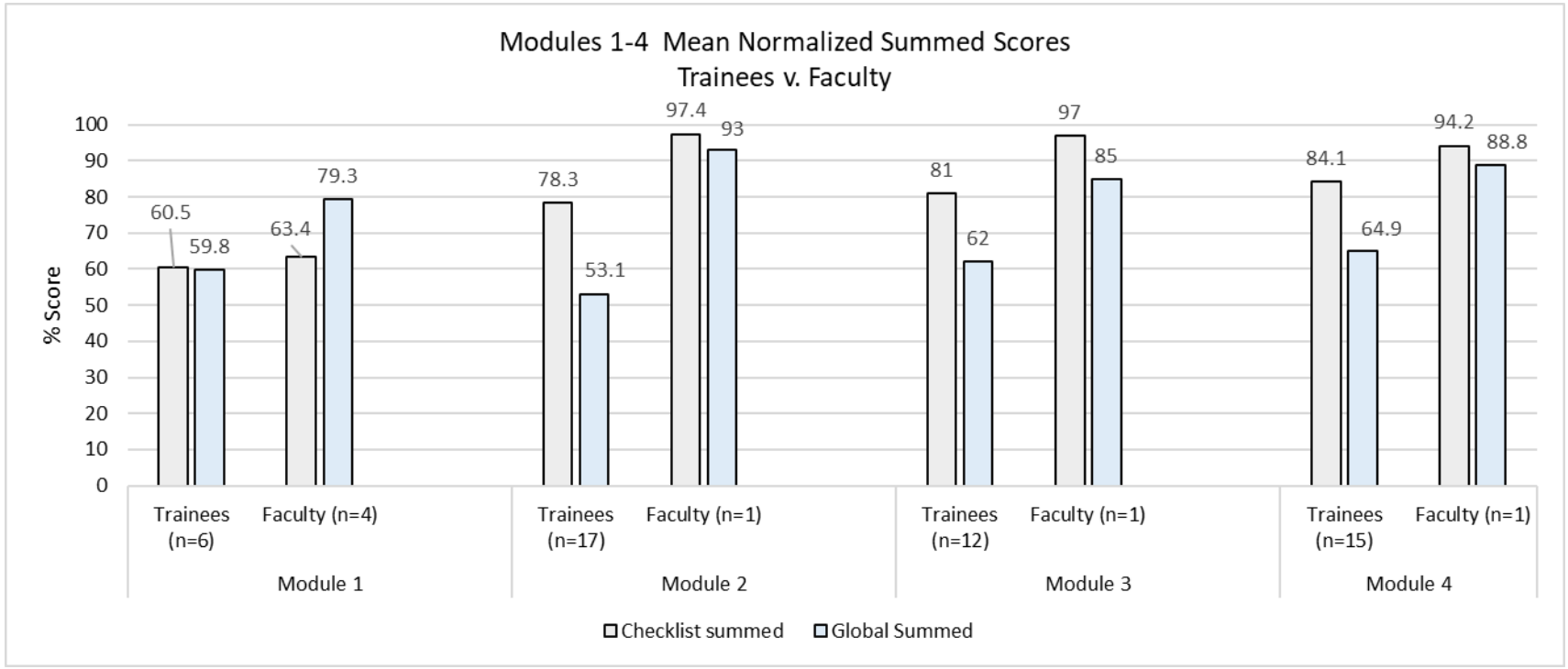
Description: Modules' Mean performance scores for Medical students' (n=10) and Residents' (n=23) performances using Kruskal-Wallis (non-normal distribution confirmed). Caveat- not all trainees, and not all raters completed each module so these are not fully linked. Because of this, independent samples analysis was performed. *Scoring:* Because all checklists had different number of items, Checklist Summed and Total Summed scores were normalized (out of 100%). Global items all max score=5, max Global sum=25

Summary

All scores were able to discriminate learners' scores across modules, exception Final rating. Interestingly, Modules were rolled out in sequence, and are associated with a positive trend across modules and time. Effect sizes are relatively low, though, which likely are caused by small sample size, which likely increased variability (SD) and decreased effect size. I suspect that effect sizes will improve as we increase the sample size.

Previous findings viewed in alternative way

Figure 1. Summary of trainees and faculty performance scores across modules



Internal Structure: Comparison of Program's Combined Ratings across Sites

Description: Comparison of performances ratings across 3 sites (Mbingo, Soddo, and UM) using Kruskal-Wallis test to ensure no statistical differences in scoring that would support generalizability of scores across sites

Table 4. Comparison of performance scores across 3 sites

item	Global Domain	Mbingo (n=13) Mean (*SD*) All Raters	Soddo (n=10) Mean (*SD*) All Raters	UM (n=10) Mean (SD) All Raters	P-value
–	Checklist SUMMED*	78.36 (16.2)	72.07 (17.2)	67.04 (20.83)	.001
1	Respect for Tissue	3.33 (.87)	3.17 (.80)	2.73 (1.03)	.001
2	Economy of Time and Motion	3.11 (.95)	2.79 (.85)	2.20 (1.13)	.001
3	Instrument Handling	3.17 (.93)	2.86 (.81)	2.17 (1.15)	.001
4	Flow of Operation	3.50 (.90)	3.23 (.78)	2.58 (1.12)	.001
5	Overall Performance	3.11 (.88)	2.81 (.78)	2.17 (1.02)	.001
–	GLOBAL SUMMED	16.21 (4.04)	14.85 (3.42)	11.85 (4.82)	.001
–	**Normalized Global Summed**	64.82 (16.17)	59.38 (13.70)	47.39 (19.27)	.001
–	**TOTAL SUMMED** *	71.83 (13.25)	65.93 (13.3)	57.79 (16.86)	.001
–	Final Rating (max=3.0)	2.35 (.75)	2.37 (.72)	2.46 (.60)	NS

Summary

There were statistically significant differences in performance scores across site, when we wanted to see no significant scoring differences. Likely explanation is that UM's operators are a less experienced group, comprised mostly of medical students, we might infer that this difference in operator experience at UM accounts for performance differences. Deeper analysis across Mbingo/Soddo might reveal something.

Table 5. Comparison of performance score across 2 sites

item	Global Domain	Mbingo (n=13) Mean (*SD*) All Raters	Soddo (n=10) Mean (*SD*) All Raters	**P-value**	**η2**	**Impact**
–	Checklist SUMMED*	78.36 (16.2)	72.07 (17.2)	.005	.04	small
1	Respect for Tissue	3.33 (.87)	3.17 (.80)	.002	.01	small
2	Economy of Time and Motion	3.11 (.95)	2.79 (.85)	.14	.01	small
3	Instrument Handling	3.17 (.93)	2.86 (.81)	.009	.03	small
4	Flow of Operation	3.50 (.90)	3.23 (.78)	.008	.02	small
5	Overall Performance	3.11 (.88)	2.81 (.78)	.023	.03	small
–	GLOBAL SUMMED	16.21 (4.04)	14.85 (3.42)	.007	.03	small
–	Normalized Global Summed	64.82 (16.17)	59.38 (13.70)	.008	.03	small
–	**TOTAL SUMMED ***	**71.83 (13.25)**	**65.93 (13.3)**	**.001**	**.05**	**moderate**
–	Final Rating (max=3.0)	2.35 (.75)	2.37 (.72)	.96	—	

Summary

Review of results comparing Mbingo and Soddo indicate statistical differences, with Mbingo trainees' performances scored higher than Soddo trainees' performances. Most scoring differences are associated with small effect sizes (**η2),** indicating site may have little practical impact on scoring differences. This being said, the Normalized Global summed score difference should not be ignored and should be examined further so we might identify the source of performance scoring differences. Next step is to review rating differences across site to see if scoring differences are legitimate performance-based or associated with possible rating bias.

Note: Comparison of performance ratings across the 3 rater sites (Mbingo, Soddo and UM) using Kruskal-Wallis test to potentially identify the source of scoring differences (is it rater-based or learner-based?) Findings are shown below.

Substantial rating differences seen across 3 sites, with UM having significantly lower ratings, suggesting UM raters, primarily medical students,

Table 6. Comparison of performance ratings across 3 sites

item	Global Domain	Mbingo (n=13) Mean (*SD*) All Raters	Soddo (n=10) Mean (*SD*) All Raters	UM (n=X) Mean (*SD*) All Raters	P-value	η2	Impact
–	Checklist SUMMED*	82.19 (17.26)	69.33 (13.31)	67.06 (20.23)	<.001	.14	
1	Respect for Tissue	3.44 (.80)	3.02 (.84)	2.85 (1.06)	<.001	.08	large
2	Economy of Time and Motion	3.09 (.91)	2.75 (.87)	2.45 (1.27)	<.001	.06	large
3	Instrument Handling	3.17 (.91)	2.83 (.87)	2.39 (1.23)	<.001	.09	large
4	Flow of Operation	3.50 (.85)	3.11 (.83)	2.91 (1.25)	<.001	.06	large
5	Overall Performance	3.20 (.79)	2.64 (.81)	2.45 (1.16)	<.001	.12	large
–	GLOBAL SUMMED	16.41 (3.71)	14.35 (3.62)	13.03 (5.49)	<.001	.09	large
–	Normalized Global Summed	65.63 (14.83)	57.40 (14.49)	52.12 (21.96)	<.001	.10	large
–	**TOTAL SUMMED ***	**74.20 (14.00)**	**63.52 (10.82)**	**60.12 (17.13)**	<.001	**.16**	large
–	Final Rating (max=3.0)	2.30 (.72)	2.37 (.72)	2.53 (.66)	.10	—	

rated overall, more severely than the other 2 sites. Looking only at Mbingo and Soddo ratings, similar patterns appeared, suggesting that scoring differences were seen across sites, with scoring consistently lower for UM (hawks) followed by Soddo, then Mbingo with highest scores across all domains/items (doves). This seems to suggest that rater training is critical to ensure inter-rater reliability, especially if the ALL-SAFE program scoring will be used as a whole.

SUMMARY: I'm guessing that each module, treated uniquely, did not pose a problem, but when combined and treated as normalized sums, rating differences were amplified. I would not recommend scoring as an entire program unless can ensure rating practice could be added as a module with feedback to ensure alignment and consistency.

Self v Other:

Nonparametric analysis (Independent Sample Mann Whitney U test) was done to identify rating differences across these 2 groups. There were statistically significant differences for normalized Summed Checklist scores (N-CHSUM, M_{Self}=23.00, M_{Other}=18.79,

p<.001), and normalized Total scores (NTSum, M_{Self}=73.54, M_{Other}=68.23, p= .002), suggesting that "self" raters are rating their own performance statistically higher than "other" raters, indicating that calibration might be important, OR alternatively we have operators rate their own performance under the pretense that it is one of many and not necessarily theirs (which I don't love because I think calibrating your own performance is a good exercise/ability.

Rasch analyses on following pages.

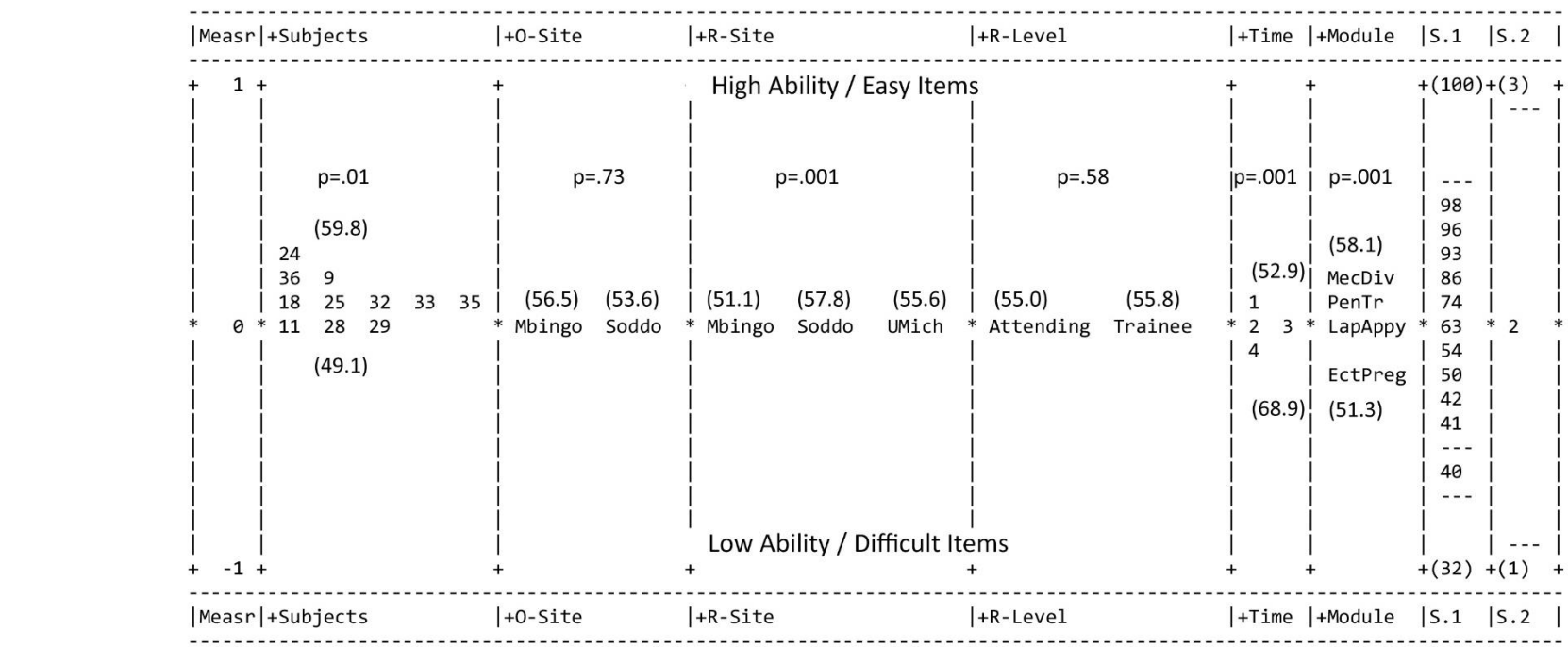
Secondary Analysis (RASCH RUN)

Description: Eleven residents completed at least 2 modules (2 modules-36.4%, 3 modules-45.5%, and 4 modules=9.1%).

A 7-facet Many Facet Rasch Model (MFRM) was used to analyze data:

Subjects/operator x operator site x Rater site x Rater Level of Experience x Time x Module x Items

Figure 2. Vertical ruler highlighting scores across different facets



Overall, the MFRM supported previous findings from classical test analyses. Findings indicated:

1. ALL-SAFE training performance measures were able to discriminate high v low performances for resident participants
2. Indicated no performance-based biases across sites
3. Indicated potential rating-based biases across sites
4. Indicated no rating biased across trainee and attending
5. Indicated that participants' scores improved over time
6. Indicated that modules were scored (high to low): Easiest→Module 4, Module 3, Module 2, Module 1→ Hardest

Point 6 was interesting, as the research team anecdotally reported that Module 3 (Penetrating Trauma) was very difficult. In spite of that, the module received fairly high scoring, suggesting that participants had improved skills enough to warrant high score, or there may have been some rating biases. Next step is to review specific biases across module and time.

Deeper bias analyses indicated no biases across site, rater experience, or time. In spite of this, I do some something notable in Table 7.

Table 7. Review of measures across modules

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Module
9300	160	58.1	70.64	.14	.01	.92	-.6	.91	-.8	1.19	4	MecDiv
8204	148	55.4	66.11	.07	.01	1.14	1.1	1.10	.9	.80	3	PenTr
7848	139	56.5	59.25	-.01	.01	1.04	.3	1.04	.4	.85	2	LapAppy
5340	104	51.3	42.14	-.20	.01	1.43	2.5	1.42	3.0	.93	1	EctPreg
7673.0	137.8	55.3	59.54	.00	.01	1.13	.9	1.12	.9			Mean (Count: 4)
1449.4	20.9	2.5	10.83	.13	.00	.19	1.2	.19	1.4			S.D.

RMSE (Model) .01 Adj S.D. .13 Separation 13.12 Reliability .99
Fixed (all same) chi-square: 626.8 d.f.: 3 significance: .00

Review of Observed Averages (OA) versus Fair-Mean Averages (FMA), I see trends that are notable:

1. Module 1 (Ectopic Pregnancy) raters were relatively forgiving (dove-like), when compared to measures from the other modules (gave OA of 51.3 and “should have” given 42.14, if using scores of other modules to calibrate)
2. Module 2 (Lap Appendectomy) raters evened out, seemed most “fair”
3. Module 3 (Penetrating Trauma) raters became more severe
4. Module 4 (Meckle’s Diverticulum) raters were most severe

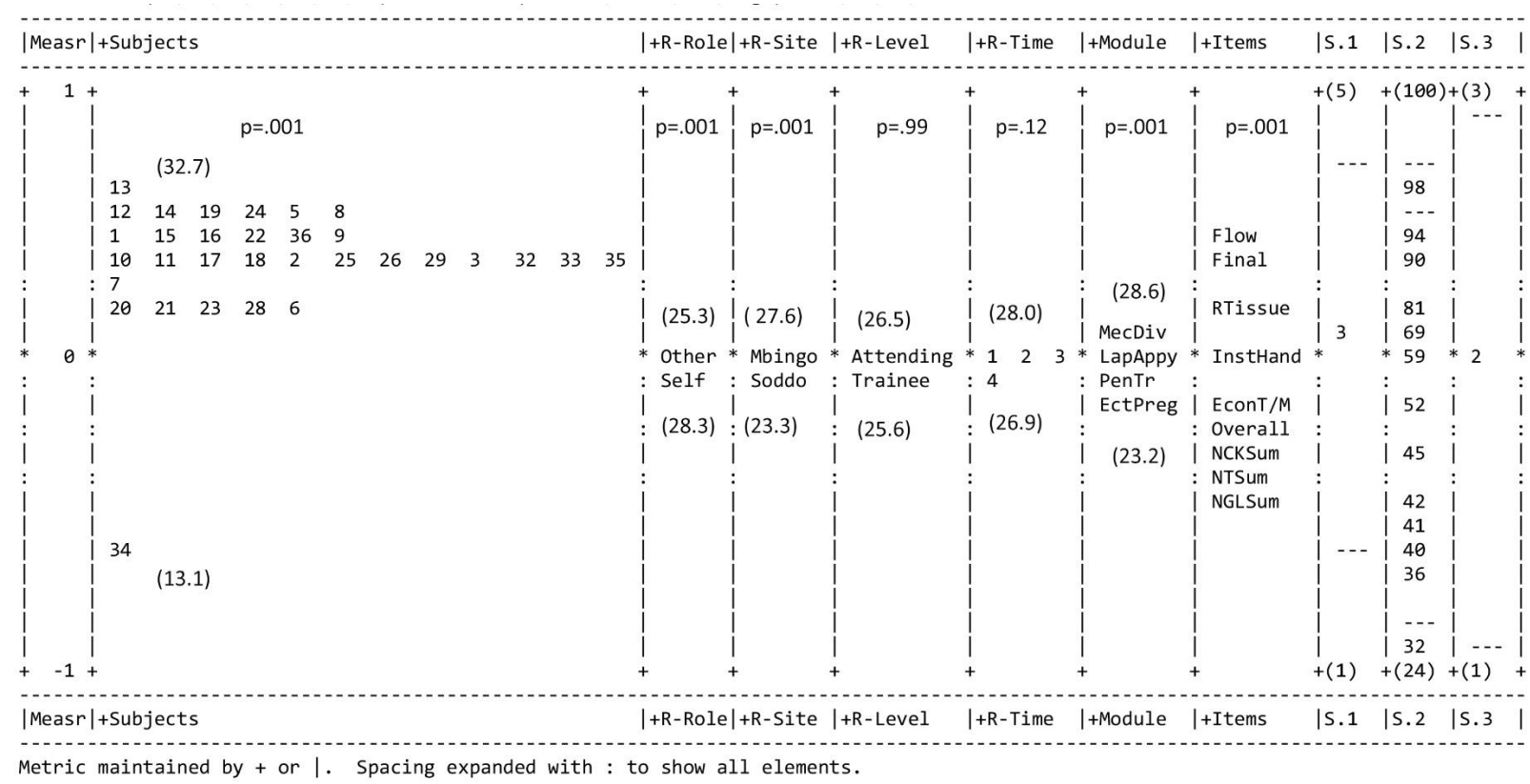
It could be that over time, raters (a mixture of medical students, residents, and faculty) may ALL have become more critical (we tend to do this as we get more confident in rating skills if no calibration). Review of rater behaviors might show something.

Rasch Analysis of Rater Behaviors

Description: 18 residents acted as *judges who rated at least 2 modules* performed by 29 residents.

A 7-facet Many Facet Rasch Model (MFRM) was used to analyze this subset of data:
Subjects/operator x Rate Role (self v other) x Rater site x Rater Level of Experience x Rater Time x Module x Items

Figure 3. Vertical ruler highlighting scores across different rater parameters



Summary

Time did not seem to impact rating behavior (thought that was possible reason for more sever ratings over time seen above). Need to review measures associated with modules to possibly ID rating biases. See Table 8, next page.

Table 8. Review of measures across modules (for raters who judged at least 2 modules)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Module
11308	396	28.6	34.06	.08	.01	.93	-.6	.76	-3.8	1.24	4	MecDiv
11736	449	26.1	33.41	.02	.01	1.06	.5	.88	-1.9	.93	2	LapAppy
11584	432	26.8	33.30	.02	.01	1.17	1.5	1.00	.0	.93	3	PenTr
15663	675	23.2	29.68	-.12	.01	1.37	3.6	1.25	4.6	.94	1	EctPreg
12572.8	488.0	26.2	32.61	.00	.01	1.13	1.3	.97	-.3			Mean (Count: 4)
1790.7	109.6	1.9	1.72	.07	.00	.16	1.5	.18	3.2			S.D.

RMSE (Model) .01 Adj S.D. .07 Separation 8.76 Reliability .99
Fixed (all same) chi-square: 359.6 d.f.: 3 significance: .00

Summary

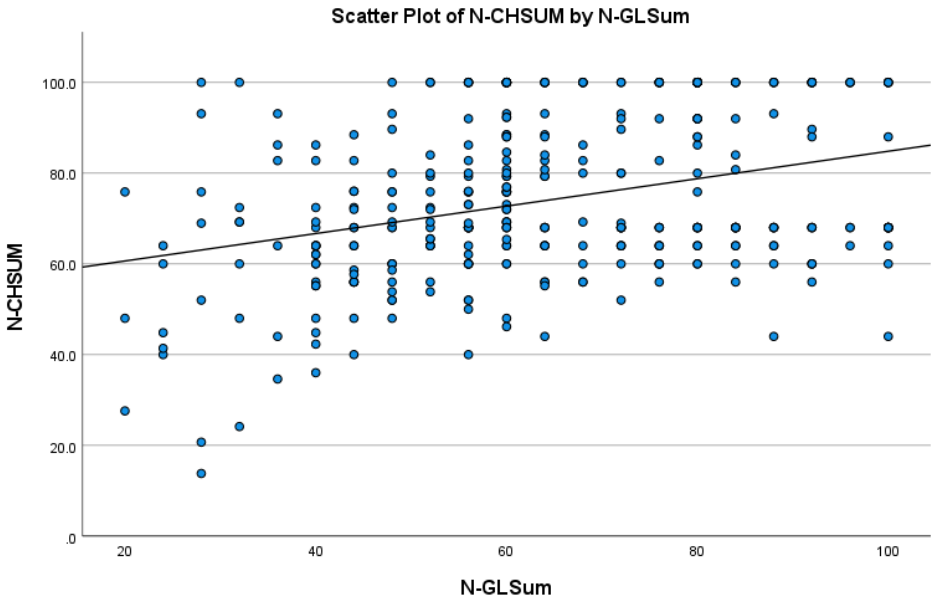
Most important trends noticed for this subgroup of raters who judged at least 2 modules:

1. Rating differences seem to be module-driven for this group of raters
2. As a whole, this group of raters seemed more severe than previous group of raters (all raters-some only rated 1 module)*
3. Ratings for Module 1 (Ectopic Pregnancy) are highly variable (note: ZStd >1.0 = “noisy”). Perhaps, lack of training
4. Ratings for Module 4 (Meckle’s Diverticulum) are highly consistent, but still a decent discriminator (Est. Discrim >1.0)

*Must not compare observed averages across 2 groups, these are thrown off in Table 8 by inclusion of checklist items (scored a 5, bringing the OA down)

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= 13 video submissions) estimated by Pearson's r.



Summary. Findings suggest a low positive correlation between normalized summed ALL-SAFE Checklist score (N-CHSUM) and normalized OSATS summed score (N-GLSum), $r(340) = .331$, $p<.001$, supporting use of the ALL-SAFE program's checklist summed score to measure performance skill.

Similarly, these summed scores (NCHSUM) correlated with the Combined summed score of the Checklist and OSATS, $r(340) = .81$, $p<.001$).

N-CHSUM and Final Rating, scored on 3-point scale (1 = Does not demonstrate competence, 2 = Borderline, 3 = Does demonstrate competence), did not correlate; $p= .46$.