2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
ASEE
Paper ID #38286

Validity evidence for measures of statistical reasoning and statistical self-efficacy with engineering students

Dr. Todd M. Fernandez, Georgia Institute of Technology

Todd is a lecturer in the Wallace H. Coulter Department of Biomedical Engineering at Georgia Institute of Technology. His research interests are engineering students beliefs about knowledge and education and how those beliefs interact with the engineering

David S. Ancalle, Department of Civil and Environmental Engineering, Kennesaw State University

David S. Ancalle is a Lecturer in the Department of Civil and Environmental Engineering at Kennesaw State University, and a PhD student in the Woodruff School of Mechanical Engineering at Georgia Institute of Technology. Ancalle earned a B.S. from the University of Puerto Rico at Mayaguez and a M.S. from the University of Illinois at Urbana-Champaign, both in civil engineering. He has a passion for teaching undergraduate engineering courses, which has driven his teaching career for the past six years. He recently began working in the area of Engineering Education and plans to continue this path after completing his graduate studies.

Validity evidence for measures of statistical reasoning and statistical self-efficacy with engineering students

1 Introduction

In this research paper, we re-evaluate structural aspects of validity for two instruments, the Current Statistics Self-Efficacy (CSSE) scale and the Statistical Reasoning Assessment (SRA) [1, 2]. The CSSE is a self-report measure of statistics self-efficacy while the SRA is a scored and criterion-based assessment of statistical reasoning skills and misconceptions. Both instruments were developed by statistics education researchers and have been consistently used to measure learning and interventions in collegiate statistics education. Our re-evaluation is part of a broader study of the effect of using a reflection-based homework grading system in a biomedical engineering statistics course [3, 4].

Prior to using scores from the two instruments to make claims in that broader study, it is a best practice to evaluate the behavior of the instruments in our use case [5, 6]. To our knowledge, this is the first use of these instruments specific to engineering students. Our justification is that the volume and level of mathematical education engineering students receive is different from that represented in a general population of undergraduates - as in prior uses of both instruments. Because the scores from any instrument are a property of *their context and use* as opposed to a property of an *instrument itself* we see such ongoing evaluation efforts as especially useful given that our population may be different [5, 7]. Substantiating that the instruments work as expected using data from our study supports relying on the scores to make claims about the homework intervention we are interested in. While re-validation is often included as a secondary component of reporting on results from instrument use, we believe reporting validation efforts separately has two benefits. First, evaluating the instrument in a separate publication supports the need for the introduction of high-quality measurement tools for engineering education as a field by making such evaluation more easily identifiable [5, 8]. Second, detaching evaluation of *a measurement's quality* from evaluation of *a measurement's results* enables a more targeted discussion of each. The detachment also separates information which may be of interest to different audiences (e.g., The score results are likely more of greater interest to instructors than the re-evaluation and vice versa with researchers).

The specific purpose of this paper is to evaluate three internal structure claims for each instrument: That (1) they *fit* established models of good measurement (e.g., consistent with prior research), (2) they are a *reliable* measure of that construct (i.e., stable across time), and (3) they are *fair* and unaffected by construct-irrelevant sources of variance (i.e., measure statistical learning; not predict gender). Other aspects of validity (e.g., content, substantive, and external validity) are also important to a holistic argument that our use is valid. However, in our case, those other claims are better established through non-empirical methods such as reviews of literature on the instrument's development, which occurred as part of the instruments' development, or evaluation of the measured constructs against the learning objectives of our course. To evaluate evidence supporting our three claims for each instrument, we organized the paper around the following research questions, with the related claim in italics:

1. How well does our data fit with prior CSSE results and characteristics of good measurement?
    (a) How well do our data align with the established factor structure? (*fit*)
    (b) How well do our data show appropriate measurement range? (*reliable*)
    (c) To what extent do demographic variables affect fit and range with our population? (*fair*)
2. To what extent do our data support scoring SRA as a single construct based on correct answers?
    (a) How well does that scoring method fit a theorized model of good measurement? (*fit*)
    (b) To what extent do items provide a useful range of difficulty? (*fit*)
    (c) To what extent do person characteristics indicate reliability of score separation? (*reliable*)
    (d) To what extent do demographic variables affect fit and score in our data? (*fair*)

2 Methods

2.1 Instruments

The Current Statistics Self-Efficacy (CSSE) scale [1] builds on established studies of students general mathematics self-efficacy to create a measure of "confidence in one's abilities to solve specific tasks related to statistics" [1]. The CSSE contains 14 self-report items with a one-sided, six-point (*no confidence at all* to *complete confidence*) Likert-like scale. The original developers established that the CSSE has a reliable uni-dimensional structure and is stable over time. They also showed that CSSE scores have a positive correlation with increasing statistics education, correlate with course performance, and have a significant and well scaled relationship with generalized mathematics self-efficacy. Later uses of CSSE consistently show similar results and good correlation with similar instruments [9, 10]. In all prior work, CSSE scores have been calculated as the sum of all responses to all items. Confirming that this approach is credible with our population is our primary goal.

The Statistical Reasoning Assessment (SRA) [2] is a multiple choice test contains 20 items. Each item contains four to six responses items representing three types - correct, incorrect, and answers reflecting known statistical misconceptions [2]. We also added a confidence rating item to each multiple-choice item, although those are not analyzed here [11]. The use of misconception response options is similar to concept inventories, an increasingly common form of assessment in engineering education [12]. The SRA was created to addresses two challenges in existing assessment methods. First, separating statistical reasoning, especially related to realistic data, from general mathematical reasoning or ability to perform computations. Second,including identified misconceptions about statistical concepts as part of measuring that understanding and reasoning [2, 13, 14, 15]. As with the CSSE, the SRA has seen continued use (e.g., [16]). The original developer's work establishes that the SRA is a "valid and reliable instrument" with good test-retest reliability and aligning with existing measures of math related reasoning in a cross-cultural study [2]. However, the developers note that mathematical measures of fit were relatively low, which is of key relevance to our study. The developers attribute low fit measures to misconceptions often being stable across education, and because each item measures a potentially distinct concept/misconception, which may be not coalesce into a generalized performance trait. Confirming the presence or absence of those behaviors in our SRA data is our primary goal because they affect how we calculate and use the scores.

2.2 Study Context and Data Collection

As noted, the data for this study is part of a broader, IRB approved, study of homework design in an undergraduate engineering statistics course during the Spring 2022 semester. The study occurred at a large, public, R1 University in the Southern United States. The university is generally considered very selective with an acceptance rate of around 20% and an average composite SAT score of 1465. The study occurred in the university's Biomedical Engineering Department (BME) - which is representative of the university in all aspects except gender balance (54% female in BME as opposed to 60% male overall). The undergraduate statistics course is a BME degree requirement, typically taken in the sophomore year, and was taught as 2 sections. The course was intended to be taught in person, but moved to an online format for the majority of the semester due to a household accident the instructor suffered during the second week of classes. The course content is generally similar to other undergraduate engineering statistics course excepting a focus on biomedical data and applications.

Data from the instruments were collected using a pre-post design using an electronic survey whose responses were combined with course exam grades after the semester's completion. All students were asked to complete a pre-survey (101 responses, 123 students, 82% participation) during the first week of the course and a post-survey (104 responses, 116 students, 90% participation) distributed in the final instructional week of the semester. The pre-survey contained a consent form, the CSSE, the SRA, and a set of demographics questions. The post-survey contained all of the pre-survey components except

demographics questions, including a re-consent. Students who completed both surveys had their lowest homework grade dropped whether or not they consented to the use of their data in the research study. Given the reward for participation, we engaged in a data cleaning process to remove low effort or incomplete responses (e.g., blank, or all the same answer) as well as responses that provided data but declined to consent to research participation [17]. Our final data set contains 196 SRA responses (99 pre and 97 post) and 183 CSSE responses (85 pre, 98 post). Because appropriate samples sizes are specific to the analyses we perform, we discuss sufficiency throughout the remainder of the paper.

2.3 Analytic Methods

We used two analytical techniques - confirmatory factor analysis (CFA, a form of factor analysis) and Rasch analysis (Rasch, a form of item response theory). In-depth descriptions and prior examples of these techniques in engineering education can be found in literature for CFA and for Rasch [18, 19]. In this section we focus on our implementation of each method and the criteria we applied to interpret the results. All analysis was performed using R Statistical Software (v4.2.0) [20]. We primarily used the Lavaan package (v0.6.12) [21] for CFA analysis and the eRm package (v1.0-2) [22] for Rasch analysis alongside related data preparation and data visualization packages.

2.3.1 Confirmatory Factor Analysis

We used CFA to evaluate Research Question 1 (RQ1), which focused on the CSSE instrument, and generally report results using guidelines for reporting factor analysis work [23], [24]. CFA enabled us to evaluate the overall fit of our data (RQ 1a and 1b) against the latent structure established in prior research [1] as well as whether that latent structure was *invariant* between groups (RQ 1b and 1c). Evaluating group invariance is a tool in identifying sources of construct-irrelevant variance that affect the interpretation of results or need for normalization of scores across groups.

Following suggestions in Byrne [25] and Hancock [23] we report the following measures of overall fit: Chi square test against a null model (chi-square), Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI, also known as non-normed fit index), and Bayesian Information Criterion (BIC). We interpret those measures using the criteria for 'good' measurement listed in Table 2. We also evaluated item loading (i.e., covariance of individual items against the latent construct) to ensure that items individually contribute to the latent construct of interest. While many guidelines are available, we adopt $<.60$ as a minimum for concern about item function and values $>.90$ as a criteria for concerns about item redundancy, which can warp scores. Those criteria are motivated by sample sizes in subgroups of interest - specifically the sample size of our pre tests was 85, a sample size at which 0.60 is a practical minimum criteria [26].

We used invariance analysis to test whether CSSE scores were consistent across groups we have reason to expect might have differences in statistical self-efficacy. Often referred to as 'fairness', establishing invariance is important to establishing whether we should correct for secondary variables (e.g., gender) when using CSSE scores in our broader study [6, 18, 27]. We evaluated invariance for three binaric groups. First, gender identity[1] because earlier research suggests lower mathematics self-efficacy in female identifying students [28, 29]. Second, pre and post responses based on the expectation that self-efficacy will likely change across a semester of instruction statistics. Invariance testing allows us to confirm that changes are limited to changes in *item score*, not changes in the underlying *scoring structure*. Confirming expected changes in mean scores exists while the underlying scoring model does not adds credibility to claims about changes in self-efficacy across the semester. Third, we compared those with and without prior

[1]The pre-survey provided multiple gender identity options using a select all that apply question format. All participants responses included either a male or female selection although some selected further options (e.g., cis, trans). Therefore the invariance test compares male- and female-identifying student groups

statistics coursework. For similar reasons to the pre and post groups, we hypothesize that prior statistics education may be a source of differences and decided to confirm the consistency of the scoring model for both groups. That confirmation, again, allows us to correct for prior statistics education in the broader study if necessary.

For each group comparison, we follow typical invariance testing procedures and applied a series of incremental constraints to the base CFA model [6, 27]. We compared three models for each group: (1) Configural - which only constrains the two groups to the same structure, (2) Weak - adding a constraint for equal factor loading between groups, and (3) Strong - adding a constraint for equal 'item intercepts' (i.e., both groups have the same mean score on each item). For each group comparison, we report the fit of models 1 vs. 2 and 2 vs. 3 as separate nested chi-square tests and also report BIC values [25].

2.3.2 Rasch analysis

To evaluate the SRA instrument, we analyzed the correct answers to SRA items only[2] using Rasch. The Rasch model, which Rasch analysis is built on, is a special case of item response theory (IRT) that fits a test (i.e., all items in an instrument) to a presupposed model of 'good measurement' in a way similar to CFA fitting to a presupposed model of a latent construct [11, 18, 30]. While typical approaches to IRT provide only a separate descriptive model for each item, Rasch provides measures test level score function and measures the fit of the test to a known model of good measurement. Our analysis consisted of two parts. First, we preformed an overall analysis (RQ 2a, 2b, and 2c), similar to analyzing fit in CFA. Second, we performed a *differential function* analysis that evaluates whether individual items or the SRA test as a whole perform similarly with different groups (RQ 2d), which is similar to the CFA invariance analysis. As with the CFA analysis, we based our work on established guidelines for performing and reporting the Rasch analysis work [11, 23, 30].

The overall analysis evaluates how well our SRA instrument data fit the Rasch models' ideal model of good measurement. The model of good measurement is unidimensional, accurate across the range of abilities present in a population, and can accurately differentiate between varying ability levels. Testing the fit of our SRA data against the Rasch model of good measurement establishes the instruments' ability to provide credible scores of statistical reasoning. From the Rasch model we report two fit statistics (infit and outfit), a person-separation reliability metric, a difficulty value for each item, and ability scores for our participants. Infit values assess fit near the average score, while outfit values assess fit at score extremes. Higher infit and outfir values indicate more randomness than is useful and lower fit values less. Criteria for productive measurement of mean square fit values varies but we adopt a 0.70-1.30 range suggested for high quality or high stakes testing [30]. Person-separation reliability is a singular value the represents the overall ability of the instrument to provide score estimates that are due to test performance rather than measurement error - with .50 suggested as a minimum and .70 as a target, but with consistent warnings to not treat the value as the sole measure of test performance. Results for item difficulty and person ability reflect log odds (in logits) that an item will be answered correctly and the ability of an individual respectively with the average difficulty and average ability constrained to zero for both [30].

The differential function analysis analyzed the same groups as the CFA invariance analysis (i.e., gender, prior statistics coursework, and pre-post test) at both the item and test level [31]. Similar to CSSE, we expect to find no differences based on gender and better performance (seen as an increase in person ability or decrease in item difficulty) in post tests or those with prior statistical experience. As with invariance

[2]Personal communication and review of prior work noted that minimal testing of a generalized scoring model for SRA using modern tools was conducted during the instrument's development. Further, as noted in the SRA development article [2], the nature of misconceptions may make them poor candidates for the application of a generalized measurement model, which we can confirm through this testing

analysis, these are important properties to confirm to guide our interpretation of results in our broader study. We evaluated differential *item* function by comparing the difficulty of each item for both groups using two criteria: A z-test provides a probability of a statistically significant difference in difficulty and a scatterplot of item probabilities with a 95% confidence interval bounds provides a visualization of how item difficulty compares between two groups. To evaluate differential test function, we perform a t-test with person ability as the dependent variable and our group comparison variable as dependent variables. Each of our three grouping variables was assessed independently.

3 Results

3.1 CSSE instrument

3.1.1 Preliminary analysis

We initially evaluated that individual item behaviors show that our data behaves as expected and is reasonable to use in CFA analysis. The results are also similar to other studies using CSSE. Item means ranged from 3.02 (CSSE.6) to 4.22 (CSSE.14) - no items have means in the lowest or highest category of the response scale. Inter-item correlations ranged from 0.41 to 0.78 and are all significantly positively correlated, as shown in Figure 1. Because CSSE is a single factor instrument, the significant same direction correlations suggest that each item can contribute to score calculation. In parallel, the lack of very large item-item correlations (i.e., $>.90$) suggest that items are probably not redundant. Measures of reliability were also positive for continuing with CFA analysis. The coefficient alpha for all items was high (0.96), which suggest our data contain only one factor. Additionally, a removed-item test did not show any items whose inclusion reduced the reliability of the instrument.

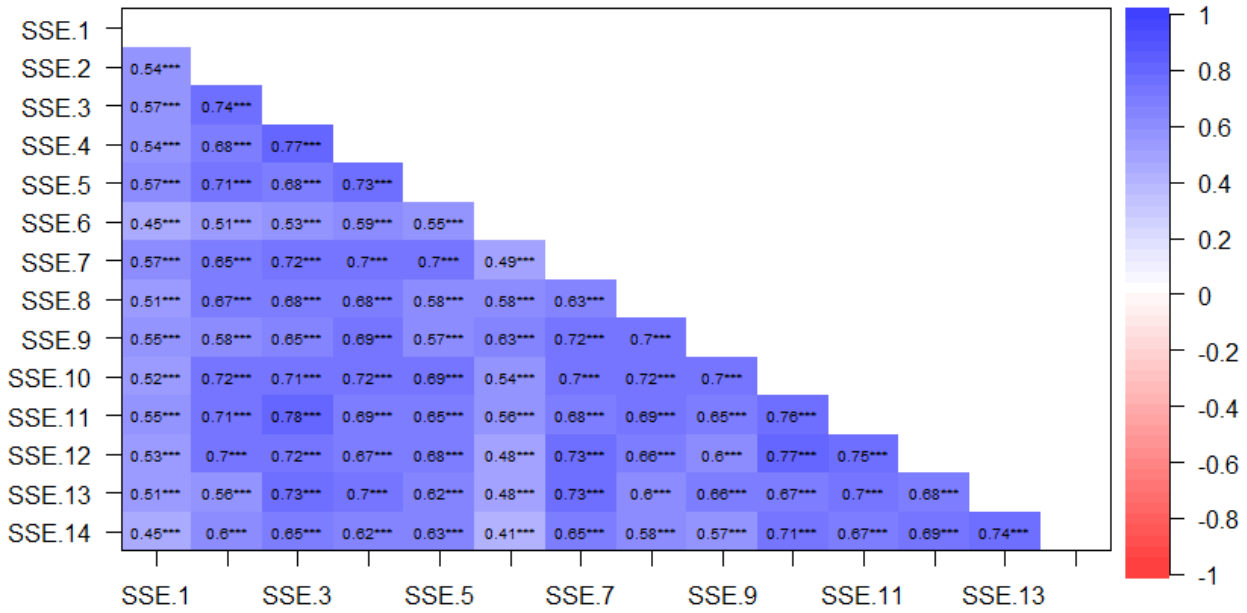

Figure 1: Plot of item-item correlations color scaled from -1 to 1 with significance demarcated using asterisks (* = $p < .05$,** = $p < .01$,*** = $p < .001$).

3.1.2 Overall model fit

We evaluated two versions of the CFA model to fully understand the latent structure of our data. Both used the single latent construct that was proposed by the CSSE developers and supported by our preliminary analysis. The first (primary) version treated individual item responses as linear, which we also compared against a (second) version treating item responses as ordered. The ordered model relaxes the assumptions

of normality and equi-interval response scale behavior to evaluate the impact of assuming that the Likert-like response scale behaves linearly.

The results (Table 1) show that both models of our data have an acceptable degree of fit with the published structure of the CSSE instrument. For the linear model, the measures of fit were at or just below criteria for acceptable fit stated in Table 2. TLI and chi square values meet the acceptable thresholds. However, both CFI and RMSEA are both just outside the acceptable criteria. Item loadings ranged from 0.64 to 0.87 with an average of 0.80. In total, all items meet our minimum threshold ($>$.60) while none suggest concerns about item redundancy ($<$.90).

Table 1: Overall fit results for CFA model of the CSSE instrument

Factors	Variables	χ^2/df	CFI	TLI	RMSEA (90% CI)	BIC
1	Linear	2.65	0.943	0.932	0.095 (.079-.111)	7577.530
1	Ordered	2.59	0.987	0.985	0.094 (.078-.110)	N/A

Notes: Both models fitted using maximum likelihood estimation. Ordered model reports robust variant which scales fit statistics and standard errors.

Table 2: Summary of CFA Fit statistics, evaluation criteria, and interpretation notes used in this study

Fit Measure	Type	Good Fit	Acceptable Fit	Notes
Chi-square	Absolute	$\chi^2/df < 2$	$\chi^2/df < 3$	higher indicates better fit
RMSEA	Absolute	< 0.05	< 0.08	lower indicates better fit
CFI	Comparative	> 0.97	> 0.95	higher indicates better fit, some suggest 0.90 and 0.95 as criteria
TLI	Comparative	> 0.95	> 0.90	higher indicates better fit
BIC	Parsimony	Lowest value is preferred		change of $>$10 is a common guideline for strong evidence of improved parsimony

Notes: Criteria for interpretation of fit statistics drawn from [24]. Criteria, especially TLI and CFI, must be interpreted with caution when evaluating models that treat variables as ordered not linear. Such models tend to inflate fit measures because of an increase in the number of fitted parameters [32].

The results also show that ordered version of the model has somewhat better overall fit. This is a typical and mathematically logical finding. Likert-like scales are ordered categorical despite frequently being treated as linear. However, more fundamentally, fitting an ordered model also simply increases the number of parameters that are fitted, which is why ordered models have a high potential for overfitting data [32]. Both the chi-square and RMSEA fit measures are effectively the same for the ordered model. However, the CFI and TLI both increase enough to meet the criteria for good fit. Interestingly, this specific result is identified in CFA literature [32]. The most common algorithms for parameter estimation (i.e., model fitting) tend to result in inflated CFI and TLI values with ordered data. While typically considered overoptimistic in CFA analysis, chi square fit indices have been shown to be relatively less optimistic about fit when used with ordered data. Item loadings for the ordered model range from 0.67 to 0.89 with an average of 0.83.

3.1.3 Measurement invariance

The measurement invariance results showed evidence of sources of misfit in the CSSE data that our aligns with soem of our expected differences in fit. Specifically, we saw violations of invariance when comparing both pre and post tests as well as those with and without prior statistics experience. In parallel, while others

have seen gender differences in mathematics self-efficacy, our results did not support those within our sample. Summary results of the invariance testing appear in Table 3.

Table 3: CFA invariance results for comparison of CSSE data by gender, pre-post, and prior statistics groups

	Gender			Pre/Post			Previous		
Constraint	χ^2/df	BIC	p	χ^2/df	BIC	p	χ^2/df	BIC	p
Configural	2.1	7980	N/A	2.2	7605	N/A	2.1	7063	N/A
Weak	2.0	7870		2.1	7548		2.0	7008	
Strong	1.9	7770		2.4	7571	***	2.0	6972	**

Note: p reports significance of a nested chi squre difference test (* = p<.05, ** = p<.01, *** = p<.001) for pairs of models with increasingly strict constraint within that group comparison. Levels of constraint are further explained in methods section.

Our results show no evidence of gender differences in self-efficacy. Constraining factor loadings (weak) does not change the fit compared to just constraining the structure (configural). Nor does constraining the item means (strong) significantly change the fit. Further, the most constrained model is also the most parsimonious as indicated by the BIC results, suggesting that the additional parameters fitted in less constrained models do not add value (i.e., unique information) to the model.

Rather than *no* differences between the pre and post responses, our comparison shows *expected* evidence of difference. For the pre and post comparison the evidence that CSSE is not invariant is not only expected, the specific types of invariance are important to our evaluation. The results in Table 3 show that constraining factor loading did not significantly change the fit, suggesting the same scoring model is valid for both pre and post tests. However, the results show that constraining the item means (strong) to be the same on the pre and post tests *does* significantly change the fit. Evidence of that importance of that effect is reinforced via the BIC values for the models. The weak model is a more parsimonious explanation of our data, despite fitting extra parameters (i.e., two intercepts for each item as opposed to only one). This suggests, as we expected, that while there is a difference in pre-post mean self-efficacy the underlying model for calculating those scores is consistent.

Similarly, we find expected evidence types of invariance when comparing students with prior statistical preparation to those without. Like the pre/post invariance tests, we see that constraining factor loading does not change the fit compared. Similarly, constraining item means shows a significant change in chi square model fit. However, the notable difference is that, the most constrained model is the most parsimonious suggesting that the effect of prior stats is less meaningful on self-efficacy scores.

3.2 SRA Instrument

3.2.1 Overall model fit

The results of the SRA show that the items reasonably fit the Rasch analysis model of good measurement overall. While participant scores are somewhat concentrated at the high end of test, which we expect to see when including pre and post tests, the items show good fit and reasonable score separation. The primary results appear in Table 4.

Table 4: Item difficulty and fit values for all SRA items

Item information		Mean square fit	
Number	Difficulty	Infit	Outfit
SRA.1	-0.78	0.92	1.06
SRA.2	-2.10	0.83	**0.67**
SRA.3	-0.19	1.01	1.01
SRA.4	-0.42	1.10	1.30
SRA.5a	-0.72	1.02	0.96
SRA.5b	0.06	0.82	0.78
SRA.8	-1.84	0.77	**0.61**
SRA.9	-1.84	0.84	0.71
SRA.10.1	-0.05	1.01	1.00
SRA.10.2	0.13	0.97	0.96
SRA.11	-0.94	0.91	0.97
SRA.12	-0.44	1.03	1.02
SRA.13	1.47	1.1	**1.33**
SRA.14	0.35	1.01	1.02
SRA.15	2.49	1.12	**1.62**
SRA.16	0.28	0.98	0.96
SRA.17	0.02	0.95	0.91
SRA.18	1.62	0.97	**1.39**
SRA.19	1.66	0.9	0.93
SRA.20	1.24	1.00	1.07
avg.	0	0.96	1.01
max	2.49	1.12	1.62
min	-2.1	0.77	0.61

Notes: Values rounded to two decimal places. Fit values outside of .70 to 1.30 criteria for good fit bolded. Average item difficulty is fixed to zero in model specification. Fit values are scaled to 1.

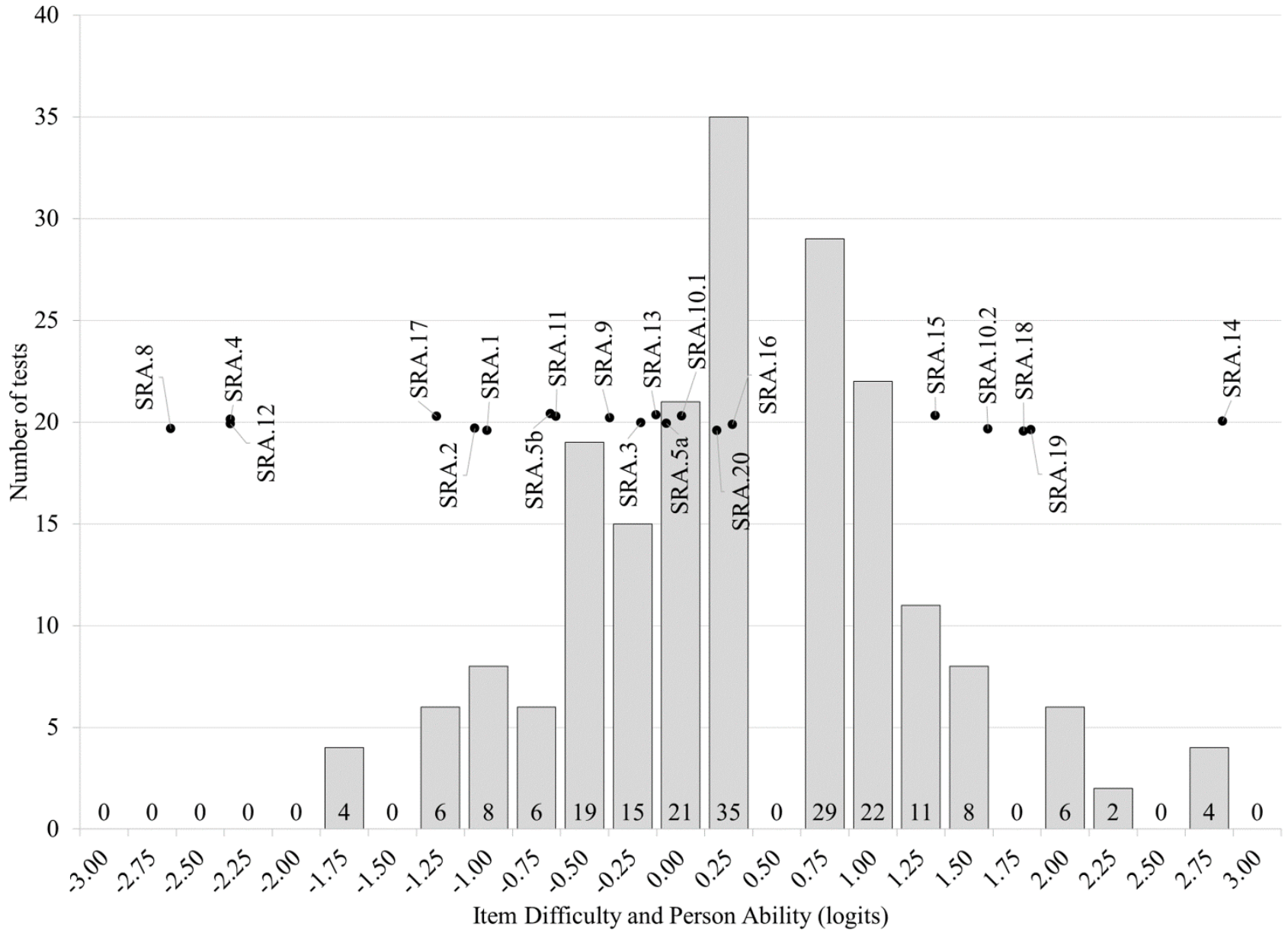

Figure 2: Comparison of item difficulties and individual scores alignment from all responses to SRA instrument. Item difficulties located and jittered on Y axis to increase interpret-ability only - item difficulty has no Y value

All 21 items had mean square infit values within the range suggested for high quality measurement - ranging from 0.77 to 1.12 against an acceptable range of 0.70 to 1.30 (avg. 0.96). Further, 16 of 21 items (min 0.61, max 1.62, avg. 1.01) also had acceptable outfit values as well. Figure 3 presents the information in Table 4 visually to highlight the relationship of fit and difficulty, as well as the overall alignment of infit and outfit values for each item.

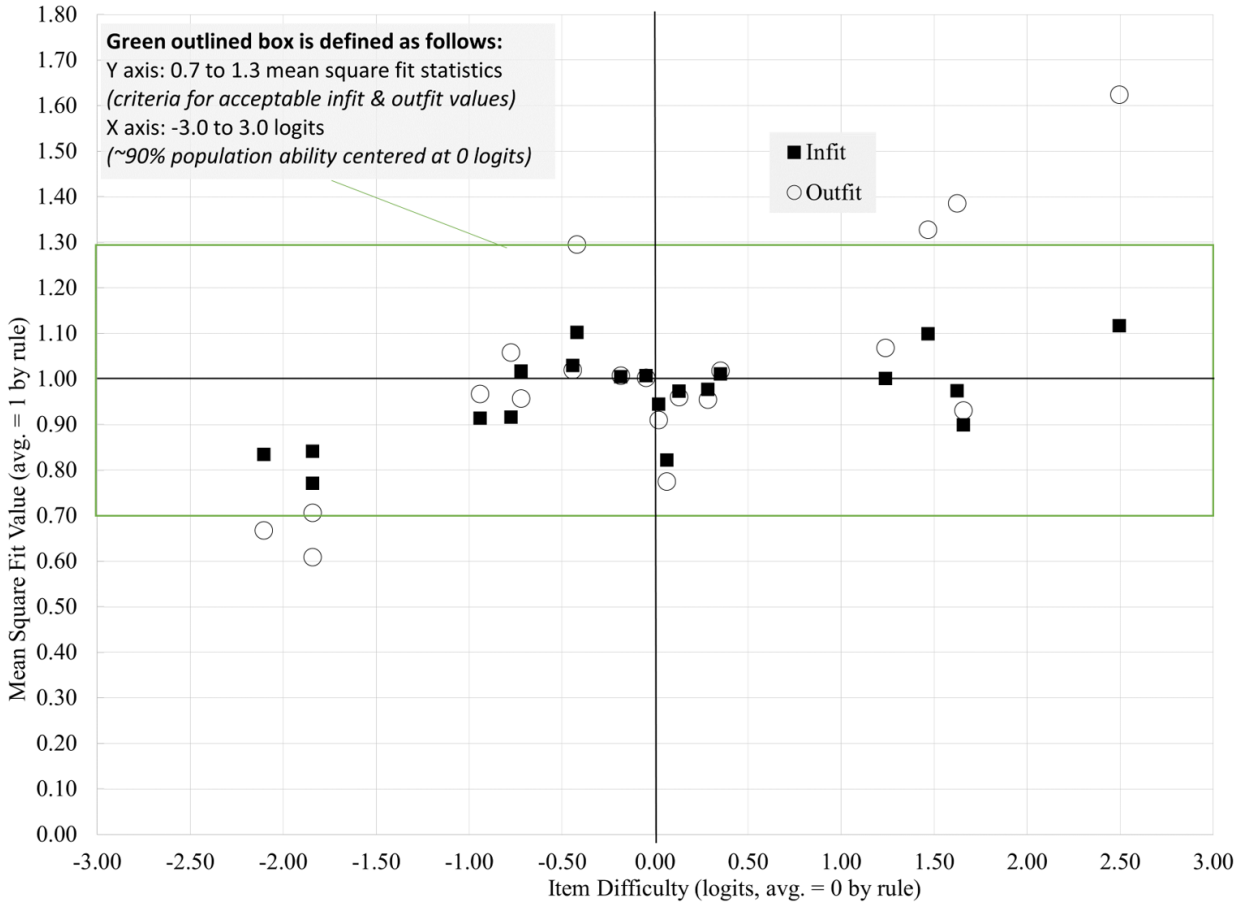

Figure 3: Comparison of mean square infit and outfit values against item difficulty for all SRA items

Of the five items outside the acceptable range, two items were below and three items above our criteria. The two low outfit items are also the two easiest items for our participants. Low outfit values do not show that items are poor measures of the construct. Instead, low outfit values indicate that they do not add information to estimate participant's ability score because there is little randomness in the outcome for particularly high or low scores - likely because almost all students get those answers correct.

The high outfit items were three of the four hardest items on the test. Two of the three high outfit (items 13 and 18) also cover concepts that not emphasized in the course used in our study (i.e., abstract probability and combinatorial reasoning respectively). Those items were retained as reference points to compare organic learning against specific course learning objective achievement. The third (item 15), with high outfit was the hardest on the SRA for our participants and involves a visual interpretation of a graph. We hypothesize some element of subjectivity in answers may be a cause here. However, all three high outfit items have reasonable infit values, suggesting they predict the majority of the sample well. Overall, the fit values show acceptable fit of our SRA data to the Rasch measurement model, although a few items that may be slightly too hard or too easy for our population.

While fit may be acceptable, the results suggest concern about score estimation and score separation using the SRA with our population. Again, baseline results are acceptable, but nuances of the SRA behavior appear less so. Item difficulty ranged from -2.10 to 2.49 logits, a range of 4.7 logits. The person separation reliability for our SRA was 0.61, below the suggested criteria of .7 but not low enough to suggest concern on its own. The average gap in difficulty between adjacent items, which is useful to understand how well covered the difficulty range is, was 0.24 logits.

However, there were five item pairs that have difficulties less than 0.05 logits apart. The poorly separated items were all below the average ability (i.e., 0 logits) of our sample. Items 8 and 9 both have a difficulty of -1.84 logits, which suggests they serve little independent value in determining score. Figure 2 shows a item-person comparison highlighting the alignment of the participant scores with the items. Three items (2, 8, and 9) have difficulties below the lowest ability score of any participant. In contrast, 4 tests have ability levels above the highest item. While concerning, we note that 3 of 4 tests above that *measurement ceiling* come from post tests. The only pre test above the measurement ceiling was from a participant whose post test was *also* above the ceiling. These results suggest the test may not be appropriately difficult.

3.2.2 Differential Function

To evaluate differential test function, we looked at the same three demographic variables as with the invariance analysis of the CSSE instrument. The results of the differential *test* function analysis in Table 5. The differential *item* function results appear in Figures 5 and 4. The results provide suggest some caution in scoring the SRA as a single instrument with these groups. Specifics for each of the three comparisons are discussed below.

Table 5: Evaluation of SRA differential test function by comparing mean scores of 3 subgroups of interest

Comparison		Group Means (logits)		Test for mean score difference		
Variable	Reference	Ref.	Comparison	t	df	p
Gender	Female	-.37	.15	-.02	81	
Prior Stats	No prior stats	.16	.29	-1.05	194	
Pre-Post	Pre	.35	.11	-1.9451	194	*

Note: All tests are independent two-sample t-tests. Based on our hypotheses, the gender comparison is two sided whereas the other two tests are single sided. * = p<.05, ** = p<.01, *** = p<.001

The comparison of scores by gender identified 2 items as having different difficulties between groups, but no evidence of different test function. Items 8 (p=.02) and 14 (p <.001) both had significantly different difficulties, although only item 14 fell outside the 95% confidence bands shown in Figure 4. Item 8 involves probabilities as ratios and was contextualized in gambling while item 14 involved sample variability in the context of gender and births. When looking at the whole test, there was no significant difference in mean test score (p=.97, two tailed) and the overall distributions (Figure 5) were generally similar between male and female identifying participants - although female participants were more likely to have scores at the upper or lower end of the distribution.

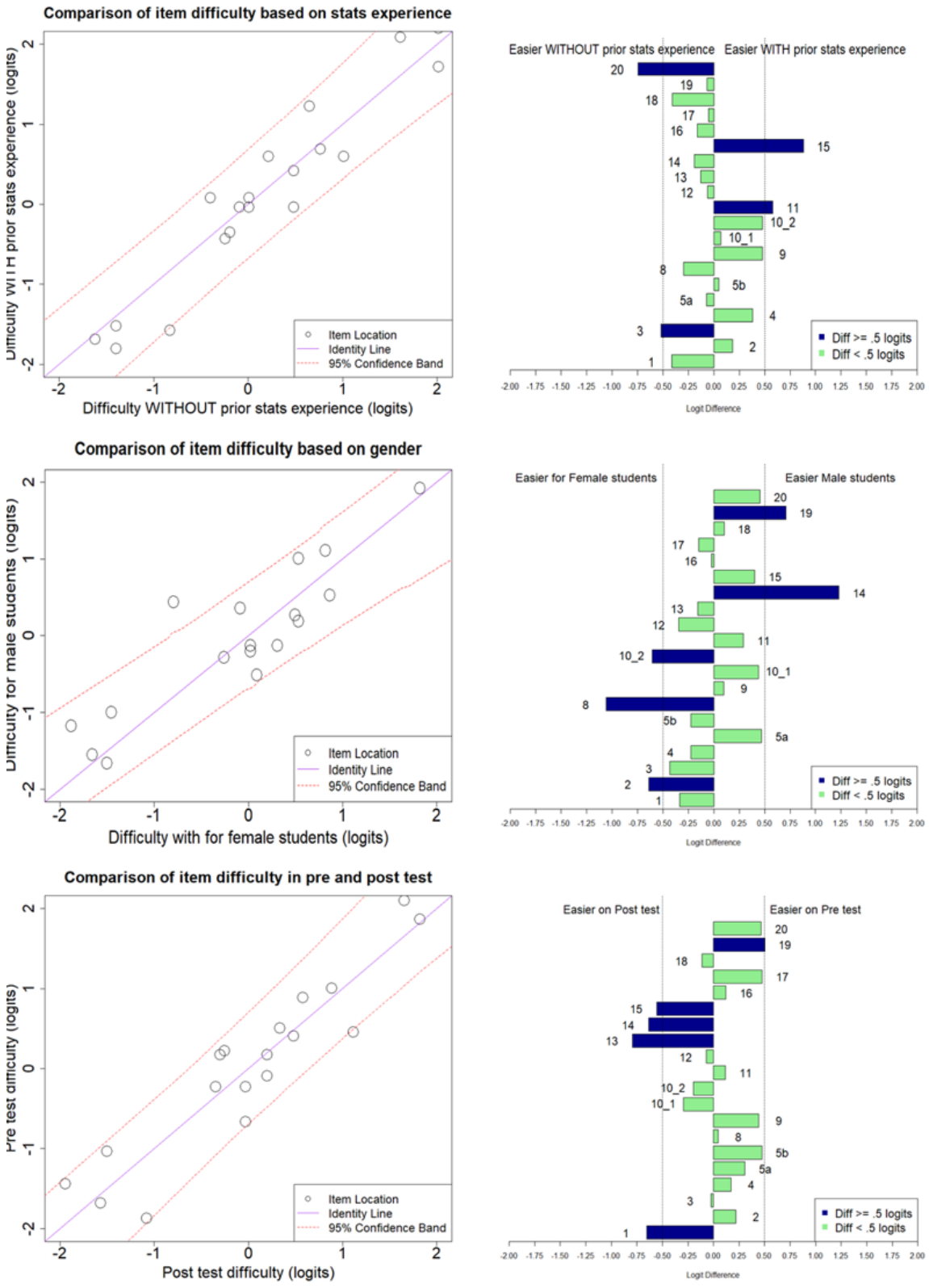

Figure 4: SRA invariance plots by group. Left column is scatter plots of item difficulties for one group on the x and the other on the y. Blue line is a y=x reference for perfect invariance. Red dotted lines are 95% confidence interval of no difference. Right column shows item by item difficulty shift in the direction of easier group. Dark blue bars show a $>.50$ logit difficulty shift - a rough criteria that is noted as potentially overly conservative [11]. Top to bottom group comparisons are prior stats, gender , and pre vs. post test

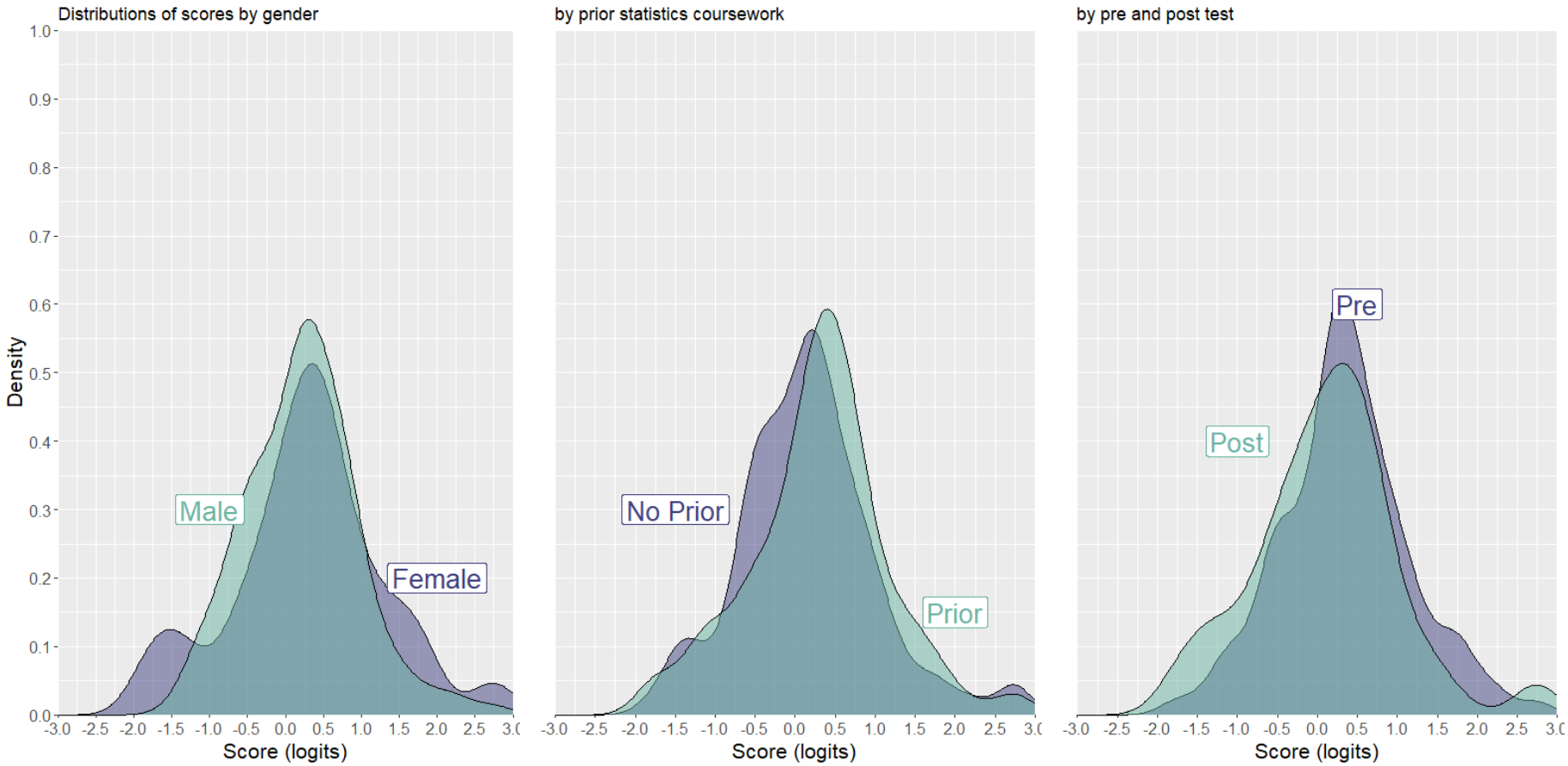

Figure 5: Comparison of Rasch estimated person score distribution by group to evaluate differential test function. Left to right gender, prior stats, pre-post group comparisons

The comparison of participants with and without prior statistics experience also show minimal evidence of differential item and test function - although the specific results were more surprising. Only item 20, which invokes rolled dice as a way to ask about combinatorial reasoning, shows differential item function. The difference is both significant (p=.02) and outside of the 95% confidence band in 4. However, that item is significantly easier for students *without* prior statistics experience, which is the opposite of what we would expect, unless the underlying reason for an incorrect answer is a misconception developed in prior statistics education. For the test overall, there is not a significant difference (p=.15) in mean scores between those with and without statistics experience - although the no prior experience group appears shifted towards lower scores. This result is different from what we expected, and from the results found for CSSE.

Finally, we compared the pre and post tests, expecting to see items become easier and scores becoming higher for post test students. We found three items were significantly easier on the post test. Item 1 (p=.04) is about selecting an appropriate average. Item 13 (p=.02, outside the 95% bounds) is about using combinatorial reasoning. Item 14 (p=.03) is about sample variability. At the overall test level, we tested for a significant increase in scores on the post test and found non (p=.97). In fact, the mean pre test scores (0.35 logits) where above the post test scores (0.11). We return to this point in the discussion.

4 Discussion and Conclusion

We use this section to summarize and further interpret our findings as they relate to how we can, and cannot, use the CSSE and SRA scores in our broader study. We separate those discussions into sections by instrument, and end by commenting on the implications for engineering education. To reiterate our standpoint on instrument validity, our results are specific to how the instruments function in our study. What limited information our results provide about each instrument *overall* is anchored in the context of prior uses, developers' intent, and prior developers' evaluative efforts - a point we address in each section.

4.1 Summarizing our use of CSSE

Our results support adopting the original scoring method (single factor, all item responses summed) to calculate a statistics self-efficacy score, as opposed to defining a new model or rejecting the model based on slightly lower fit. While higher fit values would be considered optimal, our use case involves evaluating a new pedagogy with an existing instrument. That use is low stakes and involves ongoing as opposed to new validation work. Both points that suggest that lower fit values may be acceptable [5]. Further, prior results and our secondary analyses provide credible explanations for sources and types of misfit. We see the logical explanations for reduced fit, more than fit values themselves, key evidence supporting our use of CSSE scores as credible.

In comparison to the original development work, we analyzed a similar sample size (183 vs. 140) and similar pre-post test design [1]. Our results similarly supported a single factor model ($alpha_{current} = .96$ vs. $alpha_{original} = .975$). Further, while the original developers performed exploratory as opposed confirmatory factor analysis, item loading's were also very similar (0.640 to 0.868 vs. 0.56 to 0.81). These results suggest that, at minimum, our data behaves very similarly to the developers' original study. Given that CSSE has seen ongoing use, the similarity is useful for contextualizing the results. Two additional properties not tested in the original work support the scoring method. First, we saw no evidence of ceiling or floor effects in item or score distributions. Second, we see the lack of change in fit from treating responses as ordered as support. The scoring method, sum all items, presumes equi-interval data, and we saw no improvement in fit by relaxing that assumption.

While the original developers did not test invariance, they did report other analyses we can compare to [1]. The developers found a significant increase in self-efficacy between the pre and post test. That aligns with

our result showing differences in item means between pre and the post responses. They evaluated the factor structure of the pre and the post tests separately using exploratory factor analysis. They found similar item loading, minimal change in variance explained (5%, not tested for significance), and similar support for a single factor structure. These results, similarly, align with our results that constraining pre and post responses to an equivalent scoring model did not change fit. For our broader study, demonstrating that the measurement model is the same between the pre and post test is important to evaluating change in self-efficacy across the semester. We expect *construct-relevant* variance when comparing pre and post semester scores, but need that variance to be limited to scores themselves and not the scoring model. In parallel, the results suggest we do not need to correct for a general gendered effect on self-efficacy, which prior work suggests as a *construct-irrelevant* source of variance[28]. These results also highlight the effect of fitting a single CFA model to a sample with two discrete populations (pre and post tests in our case) can negatively impact fit because of assumption of item normality that is inherent in CFA.

4.2 Summarizing our use of SRA

The results for the SRA show more caution in adopting the scoring model we evaluated in this paper. Our method of scoring made two simplifications to the SRA: (1) Treating items as correct or incorrect and (2) treating all items as measures of a single *correct statistical reasoning* construct. Those simplifications are different from the model proposed by the developers, who subdivided general statistical reasoning into 8 component skills, and scored for misconceptions as a specific type of incorrect answer. The authors specifically note that instruments which use specific misconceptions as a wrong answer choice often behave differently than traditional dichotomous scored items with a single correct, and set of generally incorrect answer choices. They link that behavior to the nature of misconceptions, especially their tendency to be stable over time. While overall the Rasch Model of good measurement fits our SRA data well, aspects of item difficulty, score distribution and differential function suggest that does not function well as a scoring method. The results have motivated us to pursue a different scoring approach.

Using our single construct scoring approach, measures of fit were acceptable. Infit values all met accepted criteria, suggesting typical scores are defensible. Further, excepting the most difficult items, the outfit results show that a broad range of scores are also predictable. As noted in the results, the person separation index was below accepted thresholds. That measure estimates whether the distribution of item difficulty in the overall instrument is sensitive enough to differentiate between people of different abilities [33]. However, in cases where other indications of model fit and function are positive, that measure alone is not a reason to reject a scoring model.

However, the difficulty of the items limits the interpretative value of those predictable scores. The results show several cases of items with near equivalent difficulty. While those items focus on different component skills, our intended approach focuses on one score of an overall construct. In those cases, multiple near equivalent items mean that a simple approach to score calculation (i.e., summing the raw number of correct answers) will likely warp scores. While that problem can be solved by using the participant ability levels calculated by the Rasch model, doing so has other problems. Three items had difficulties below the lowest scoring participant - meaning they add no value to estimating ability levels. Similarly, the grouping of item difficulties means large areas of the ability spectrum have little coverage, meaning scores in them are less accurate. The cumulative effect is visible in Figure 2 and explains the low separation reliability. For scores below the mean of our population, 12 items contribute to score calculation and separation. In contrast, for abilities above the mean, only 4 items can separate between different scores, with items 18 and 19 doing so redundantly. Again, these results speak to our scoring method not necessarily speak to the SRA itself. We saw further reasons for concern in the differential function test, which we believe reflects the nature of misconceptions the SRA developers noted. Primarily, we were surprised by the results showing gendered item function and no meaningful difference in ability levels between pre and post responses. We have no

basis for expecting a difference in female and male performance on specific items, however the results show two items with such a difference. Most interestingly, the difference occurred specifically on the easiest item (item 8), which was easier for female students, and the hardest item (item 14), which was easier for male students. We have no explanation for this, only a hypothesis that different test taking behaviors may contribute. For the pre and post comparison, we expected to see the test and items become easier in the post responses. Three of twenty items did exactly that. However, the significance threshold in differential item testing is only one aspect and is limited to identifying specific items of concern. When looking at all item difficulties, items were approximately equally likely to be harder on the pre test or the post test. That effect is apparent in the differential test function analysis with pre test ability scores higher than post test. We hypothesize two potential causes - the first is the potential for low effort responses, which we had attempted to address through data cleaning as noted in the methods. The second is the presence of specific misconceptions the developers warn about - which may have been created, reinforced, or retained, through the course. Overall, our results do not support either of our scoring simplifications. As described in the methods section, the development of SRA did not involve partial credit modeling or modern scale-level analysis [2]. The developers focused on identifying the presence of a broad set of correct reasoning skills and related misconceptions, rather than a singular score [Personal Communication, 2022]. Our results suggest that both the broad range of skills, and modeling of misconceptions as different from general incorrect answers have value to interpreting SRA scores. In future work, we plan to use regression techniques to introduce both of those considerations to evaluate the SRA data as part of our broader study.

4.3 Implications for engineering educators

For the field, we see two primary findings useful to other scholars. First, is that these instruments do behave in rational ways and are likely useful for understanding engineering students' statistical learning. The CSSE, specifically, works well with engineering students and is a useful tool to evaluate a specific sub area of self-efficacy. Given the significant math focus of many engineering programs, an instrument that separates out statistics self-efficacy from general mathematics self-efficacy may be useful. The SRA results do not support our approach to generating a single score, but do show strong evidence that the individual items behave as expected and that the instrument has generally valid behavior for capturing data about engineering students' statistical reasoning. In those ways, these instruments are a useful contribution to the body of measurement tools within engineering education that have data supporting their use specifically with our student population.

The results also highlight the ways in which tools like CFA and IRT can be extended to draw deep insights about instruments. Such insights about an instrument are fundamental to appropriately interpreting and relying on the scores that they provide for a specific purpose [5]. While testing the fit of a theorized factor structure is on its own important, invariance analysis can provide additional information about instruments' validity. For the CSSE instrument - invariance analysis explicitly demonstrated that the latent model of the instrument was consistent across our pre and post test. Without showing that, results that treat the model as the same would be less credible. Similarly, the Rasch model itself, as opposed to classical test theory or IRT, provided a way to test the validity and reliability of simplifications to the instrument's proposed scoring system. Then, differential function analyses also provided information about how the scores and their meaning changed. More than the Rasch model results, that comparison across groups demonstrated that our proposed scoring model was not defensible and gave strong indications as to why. Whether CFA or Rasch models, information gleaned from invariance and differential function analysies are important. The results can establish, or challenge, assumptions that scores mean the same thing across time or subgroups as well as that scores change in expected ways. While meaning is often assumed to be static in many types of measurements that engineers perform in technical work, for educational instruments that assumption

should be, and can be, validated. Both are critical to supporting studies, like ours, about the impact of a novel educational intervention.

References

[1] S. J. Finney and G. Schraw, "Self-efficacy beliefs in college statistics courses," *Contemp. Educ. Psychol.*, vol. 28, no. 2, pp. 161–186, Apr. 2003.

[2] J. B. Garfield, "Assessing statistical reasoning," *J. Educ. Behav. Stat.*, vol. 2, no. 1, pp. 22–38, May 2003.

[3] T. A. Wood, D. D. Nale, and R. K. Giles, "Closing the homework feedback loop using Dual-Submission-with-Reflection homework methodology," in *2020 ASEE Virtual Annual Conference Content Access*, Jun. 2020.

[4] C. R. Lund, "Can students Self-Generate appropriately targeted feedback on their own solutions in a Problem-Solving context?" in *2020 ASEE Virtual Annual Conference Content Access*, 2020.

[5] K. A. Douglas and Ş. Purzer, "Validity: Meaning and relevancy in assessment for engineering education research," *J. Eng. Educ.*, vol. 104, no. 2, pp. 108–118, Apr. 2015.

[6] T. M. Haladyna and S. M. Downing, "Construct-irrelevant variance in high-stakes testing," *Educ. Meas. Issu. Pr.*, vol. 23, no. 1, pp. 17–27, Oct. 2005.

[7] S. Messick, "Validity of psychological assessment," *Am. Psychol.*, vol. 50, no. 9, pp. 741–749, 1995.

[8] B. M. Olds, B. M. Moskal, and R. L. Miller, "Assessment in engineering education: Evolution, approaches and future collaborations," *Journal of Engineering Education*, vol. 94, no. 1, pp. 13–25, 2013.

[9] S. Loewen, E. Lavolette, L. A. Spino, M. Papi, J. Schmidtke, S. Sterling, and D. Wolff, "Statistical literacy among applied linguists and second language acquisition researchers," *Tesol Quarterly*, vol. 48, no. 2, pp. 360–388, 2014.

[10] E. R. Walker and K. E. Brakke, "Undergraduate psychology students' efficacy and attitudes across introductory and advanced statistics courses." *Scholarship of Teaching and Learning in Psychology*, vol. 3, no. 2, p. 132, 2017.

[11] I. Paek, J. Lee, L. Stankov, and M. Wilson, "A study of confidence and accuracy using the Rasch modeling procedures," *ETS res. rep. ser.*, vol. 2008, no. 2, pp. i–25, Dec. 2008.

[12] D. Evans, G. L. Gray, S. Krause, J. Martin, C. Midkiff, B. M. Notaros, M. Pavelich, D. Rancour, T. Reed-Rhoads, P. Steif *et al.*, "Progress on concept inventory assessment tools," in *33rd Annual Frontiers in Education, 2003. FIE 2003.*, vol. 1. IEEE, 2003, pp. T4G–1.

[13] J. Garfield and A. Ahlgren, "Difficulties in learning basic concepts in probability and statistics: Implications for research," *Journal for research in Mathematics Education*, vol. 19, no. 1, pp. 44–63, 1988.

[14] D. Kahneman, S. P. Slovic, P. Slovic, and A. Tversky, *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[15] R. C. delMas, "A comparison of mathematical and statistical reasoning," in *The challenge of developing statistical literacy, reasoning and thinking*. Springer, 2004, pp. 79–95.

[16] P. F. Tremblay, R. Gardner, and G. Heipel, "A model of the relationships among measures of affect, aptitude, and performance in introductory statistics." *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, vol. 32, no. 1, p. 40, 2000.

[17] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data." *Psychological methods*, vol. 17, no. 3, p. 437, 2012.

[18] K. A. Douglas, T. M. Fernandez, S. Purzer, M. Fosmire, and A. Van Epps, "The critical-thinking engineering information literacy test (celt): A validation study for fair use among diverse students," *The International journal of engineering education*, vol. 34, no. 4, pp. 1347–1362, 2018.

[19] K. A. Douglas, T. Fernandez, M. Fosmire, A. S. Van Epps, and S. Purzer, "Self-directed information literacy scale: A comprehensive validation study," *Journal of Engineering Education*, vol. 109, no. 4, pp. 685–703, 2020.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[21] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.

[22] R. Debelak and I. Koller, "Testing the Local Independence Assumption of the Rasch Model With Q3-Based Nonparametric Model Tests," *Applied Psychological Measurement*, 2019.

[23] G. R. Hancock, R. O. Mueller, and L. M. Stapleton, *The reviewer's guide to quantitative methods in the social sciences*. Routledge, 2010.

[24] K. Schermelleh-Engel, H. Moosbrugger, and H. Müller, "Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures," $https://www.stats.ox.ac.uk/\ snijders/mpr_Schermelleh.pdf$, $Accessed: 2023-1-13$.

[25] B. M. Byrne, *Structural equation modeling with mplus: Basic concepts, applications, and programming*, ser. Multivariate Applications Series. London, England: Routledge, Jun. 2013.

[26] J. F. Hair, *Multivariate data analysis*. Prentice Hall, 2009.

[27] F. F. Chen, K. H. Sousa, and S. G. West, "Teacher's corner: Testing measurement invariance of Second-Order factor models," *Struct. Equ. Modeling*, vol. 12, no. 3, pp. 471–492, Jul. 2005.

[28] F. Pajares, "Gender differences in mathematics Self-Efficacy beliefs," in *Gender differences in mathematics: An integrative psychological approach , (pp*, A. M. Gallagher, Ed. New York, NY, US: Cambridge University Press, xvi, 2005, vol. 351, pp. 294–315.

[29] R. A. Louis and J. M. Mistele, "The differences in scores and Self-Efficacy by student gender in mathematics and science," *International Journal of Science and Mathematics Education*, vol. 10, no. 5, pp. 1163–1190, Oct. 2012.

[30] B. D. Wright and G. N. Masters, *Rating Scale Analysis*. Pluribus Press, 1982.

[31] J. Brodersen, D. Meads, S. Kreiner, H. Thorsen, L. Doward, and S. McKenna, "Methodological aspects of differential item functioning in the rasch model," *Journal of Medical Economics*, vol. 10, no. 3, pp. 309–324, 2007.

[32] Y. Xia and Y. Yang, "RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods," *Behavior research methods*, vol. 51, pp. 409–428, 2019.

[33] B. D. Wright and G. N. Masters, *Rating scale analysis*. MESA press, 1982.