(Re)Validating Cognitive Introductory Computing Instruments

Ryan Bockmon
University of Nebraska - Lincoln
Lincoln, Nebraska
ryan.bockmon@huskers.unl.edu

Stephen Cooper
University of Nebraska - Lincoln
Lincoln, Nebraska
scooper22@unl.edu

Jonathan Gratch
Texas Woman's University
Denton, Texas
jgratch@twu.edu.

Mohsen Dorodchi
University of North Carolina at Charlotte
Charlotte, North Carolina
mdorodch@uncc.edu

ABSTRACT

Cognitive tests have been long used as a measure of student knowledge, ability, and as a predictor for success in engineering and computer science. However, these tests are not without their own problems relating to priming, difficulty (resulting in test fatigue) and time on exam. This paper discusses efforts to modify Parker et al.'s Second CS1 aptitude test (SCS1) [13] to reduce the time spent on the exam, provide greater customization to match concepts taught across three universities, and reduce redundancy of test questions all while maintaining the instrument's reliability. This instrument was modified for use on an ongoing grant investigating whether spatial abilities impact the success of students in introductory CS courses. The instrument developed in this paper is a revised shortened version of Second Computer Science 1 (SCS1) aptitude test, designated as SCS1R.

CCS CONCEPTS

• **Social and Professional topics → Student assessment**;

KEYWORDS

Assessment; CS1; Validity; Replication

ACM Reference Format:
Ryan Bockmon, Stephen Cooper, Jonathan Gratch, and Mohsen Dorodchi. 2019. (Re)Validating Cognitive Introductory Computing Instruments. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19), February 27-March 2, 2019, Minneapolis, MN, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3287324.3287372

1 INTRODUCTION

Computer science instructional intervention projects center on determining the impact that they have on student learning. To accurately measure the level of success of an intervention, researchers must have and use valid and reliable instruments. However, the field of computing has few valid and reliable assessments for pedagogical or research purposes [2].

Our goal was to identify valid and reliable instruments to collect baseline data used for our research grant investigating whether improving first year computing students' spatial skills impacts their success in computer science, as well as looking into how other factors might play a role in student success in first year computing classes. To meet these needs, the project required a valid instrument to test students' computing ability, a valid instrument to measure affective interests, and a valid instrument to test students' spatial skills. Many of the instruments available for use did not completely fit project needs.

During the summer of 2014 Cooper et al. ran a two-week summer coding workshop where they taught both a control and a treatment group of rising twelfth grade students [8]. The treatment group received a 45-minute spatial skills training session each morning in place of a review of the previous day's material. Four instruments were used to collect data. The first instrument collected data on demographics. The second instrument gathered information about students' confidence towards learning computing as well as gender roles concerning computing. The third and fourth instruments were the AP Computer Science exam [6] to test students' programming abilities and the Revised Purdue Spatial Visualization Test: Rotations (PSVT:R) to assess student spatial visual acuity [20]. At the end of the workshop, the researchers concluded that the treatment group had a greater gain between pre-test and post-test scores on the computer science instrument as well as having higher confidence with respect to their perceived programming experience.

In planning to replicate and expand on Cooper's study we ran into a few issues. One of which was the use of the AP Computer Science exam. While Cooper study was ran on a small group of students our project is being conducted across three universities, where each participating university teaches CS1 using different programming languages while the AP Computer Science exam is written for Java. This paper discuses the efforts of finding and revalidating a CS1 aptitude test for our study. While the validity of the other three instruments are of interest to us as well they fall outside the scope of this paper.

1.1 CS Aptitude Tests

Aptitude tests are a systemic means of testing a learner's abilities to perform specific tasks and react to different situations. Generally, the exam has a standardized method of administration and

SIGCSE '19, February 27-March 2, 2019, Minneapolis, MN, USA

ACM ISBN 978-1-4503-5890-3/19/02…$15.00
https://doi.org/10.1145/3287324.3287372

scoring, with results quantified and compared to other test takers. For CS, these competencies can include concepts such as logical or numerical reasoning, pattern recognition, or program simulations. As they represent a standardized exam, they can be used to measure a CS student's overall aptitude toward success in the field. However, variations within required competencies and curriculum valued or accepted at a specific location may differ resulting in incongruous matches between taught content and exam content.

Our spatial skills project is being conducted at three universities. Each participating university teaches CS1 using different programming languages, so the project required a single content instrument that could be used at our three schools. There are a few attempts to create and validate content instruments.

The most well-known CS content instrument is the AP Computer Science exam [6]. However, the AP CS asks questions in Java, which does not work for our study, given that two of our schools teach CS1 in other languages.

Tew's Foundational CS1 Assessment instrument (FCS1) [18] attempted to create a CS assessment test independent of any single programming language. FCS1 tests student ability using pseudo code written in the style of imperative languages. The validation process was conducted through a multi-step process that included an expert review panel, large-scale comparison between the FCS1 and language-dependent isomorphic tests [17, 18]. The study was conducted with 952 participants across two different universities with three different programming languages being taught: Java, Matlab, and Python. Because there was no other test validated beforehand they used the final exam scores of each participant and tested for correlations. Results showed that Java having the highest correlations with a Pearson's r = .511, $p < 0.001$. Python originally had the lowest correlation, but after splitting into two subgroups results showed that the CS-python focused section had a Pearson's r = .679, $p < 0.001$ and Media-Python had a Pearson's r = .262, $p < 0.001$. It was concluded that the FSC1 is a valid instrument [18]. This instrument met our multi-language needs. Unfortunately this instrument was unavailable for use.

The Second CS1 (SCS1) exam [13], a successor of the FCS1, was created in 2016. The study was conducted by testing a group of students (n = 183) on both the SCS1 and the FCS1 exams. The results found that there was a strong positive correlation between student scores on both the SCS1 and the FCS1 with a Pearson's correlation = 0.566, $p < 0.001$. Running a quantitative analysis using a 3 parameter Item Response Theory (IRT), the researchers indicated that both the FCS1 and the SCS1 were quite difficult. Testing reliability showed that both the FCS1 and the SCS1 failed to reach a Cronbach's alpha of a 0.65 [9] with their FCS1 having a Cronbach's alpha of a 0.53 and the SCS1 with a Cronbach's alpha of a 0.59 [13]. Like the FCS1, the SCS1 is multi-language (and was validated in Java, Python, and Matlab). The problem with using this exam as written was the length of time it takes students to complete.

The Computer Programming Aptitude Test was created by the University of Kent and designed to be free of any knowledge of programming languages. It consists of 26 questions composed of numerical problem solving, logical reasoning, attention to detail, pattern recognition and the ability to follow complex procedures [1]. To date, there has been no attempt to fully validate this test. A study has indicated that there is a correlation between scores on Computer Programming Aptitude Test and final grades [11]. The lack of formal validation and the requirement of an instrument that consisted of programming questions resulted in this instrument not being selected.

Despite concerns over the instrument's length, we decided to use the SCS1 and modify the instrument by taking a subset of the original SCS1 questions. The modification to the original SCS1 required us to establish the reliability for the new version of the test.

2 METHODS

Our data was collected in a pre-post format in introductory computing classes during the fall semester of 2017 across three universities: the University of North Carolina at Charlotte, Texas Woman's University, and the University of Nebraska - Lincoln. The instrument was available for students to take during the first 3 weeks of class as a pre-test and offered during the final 3 weeks as a post-test. 635 students took our modified SCS1R.

As part of the study, students were administered four different surveys: SCS1R to measure student computer science content knowledge, an attitudes instrument to examine the students' current attitudes and perceptions on CS, the Revised Purdue Spatial Visualization Test: Rotations (RPSVT:R) to assess student spatial visual acuity, and a demographics survey to collect information on students' background. To help reduce test fatigue we aimed to keep the time to complete all four surveys under 45 minutes. All participation was voluntary, with incentives used to encourage students to participate. All incentives were approved by each institution's IRB process and varied between institutions, consisting of either time slotted during labs, $10 gift cards, and/or extra credit.

2.1 SCS1R Design

The original SCS1 consists of 27 pseudo code questions, covering the typical CS1 topic coverage. Faculty taking the test required over an hour to complete. Students pre-flighting the test required 2 to 3 hours to complete it. The researchers wanted the test duration to be no more than 20-30 minutes to complete given that the spatial skills test would take up to 20 minutes and the demographics survey combined with the attitudes test would take about 10 minutes to finish. To keep the test length to a maximum of 20 minutes we selected 9 questions, retaining one item for each major topic covered at each of the first year computing course at the participating universities. These topics include; if statements, logic operations, variable declaration, arrays, return values, code completion, code tracing, and for and $while$ loops. When multiple questions covered a given topic, we selected the one question we judged to be the best fit for our study. In the end we retained questions 1, 3, 8, 9, 11, 22, 25, 26, and 27 from the original SCS1, now labeled as Q1 - Q9. Question 1 covered basic for loop logic. Question 3 covered infinite $while$ loops. Question 8 covered more advanced for loops having embedded loops. Question 9 covered more advanced $while$ loops with the mod operator and basic boolean logic. Question 11 covered function calls and boolean logic. Question 22 covered arrays and array adding and subtracting. Question 25 covered converting math equations to code. Question 26 covered advanced boolean logic, and question 27 covered function calls and variable declaration. The interested reader should contact the instrument authors [13] for access to the questions.

3 ANALYSIS

A Rasch Model (Item Response Theory (IRT)) was used to test reliability and validity of the SCS1R. We used Ganz Rasch, a free software program, to apply IRT [3]. Running IRT on the SCS1R we were able to calculate the difficulty level of each question, plotted each question's item characteristic curve (ICC), and tested the fit of each item. There also exist 1-par, 2-par and 3-par IRT Models. The recommend sample size to run a 1-par is N = 150 for a 10 item test and N = 750 for a 2-par and 3-par IRT [15]. Due to the sample size (N = 635) and limited access to software, the Rasch model was decided to be the best fit. The researchers intend to switch over to a 2-par or 3-par as the sample population increases.

Calculations of Cronbach's α was used to test reliability of the SCS1R. An acceptable minimum value of .70 is considered acceptable for both Cronbach's α and ordinal α in educational research [12].

We used a t-test to compare students SCS1R scores to their final letter grades. Where the final letter grades was split into two groups. A t-test was used to determine if there was a significant difference of SCS1R scores between the two groups. While a chi-square and an ANOVA tests for significance differences of scores between multiple groups, a t-test is the best fit for comparing two groups [16].

3.1 Item Response Theory

Item response theory attempts to model the relationship between an examinee's ability, Θ, and the probability, P, of the examinee correctly responding to any particular test question, *i* [5]. The Rasch Model is defined below. β is the difficulty of a given question, *i*, and θ the ability of an examinee.

$$P_i(\Theta) = \frac{e^{(\Theta - \beta_i)}}{1 + e^{(\Theta - \beta_i)}} \tag{1}$$

Abilities are calculated based on how well an examinee does on the overall test (equation 2). Equation 2 does not work for examinees who score a 100% or 0%. While you might want the examinee to get a 100% on the test, that does not assist in developing an understanding of their abilities and points to other typical explanations such as a low difficulty level of the exam. The same is true for examinees who score 0. The specific details, while beyond the scope of this paper, can be found in Baker and Kim [5].

$$\Theta_j = ln \frac{\%Correct}{1 - \%Correct} \tag{2}$$

Difficulties are calculated based on how many examinees answer a question correctly. Equation 3 displays a rough estimate on the calculation of a question's difficulty. We used conditional maximum likelihood (CML) estimation. Further details are available in [4, 5]. Similarly to equation 2, equation 3 does not work for questions that have 100% correct responses or 0% correct responses. In a similar way as not wanting an examinee to get a score of 100 or 0, having a question that examinees get right 100% or 0% of the time will not benefit the research when trying to accurately measure examinees' abilities.

$$\beta_j = ln \frac{1 - \%CorrectResponses}{\%CorrectResponses} \tag{3}$$

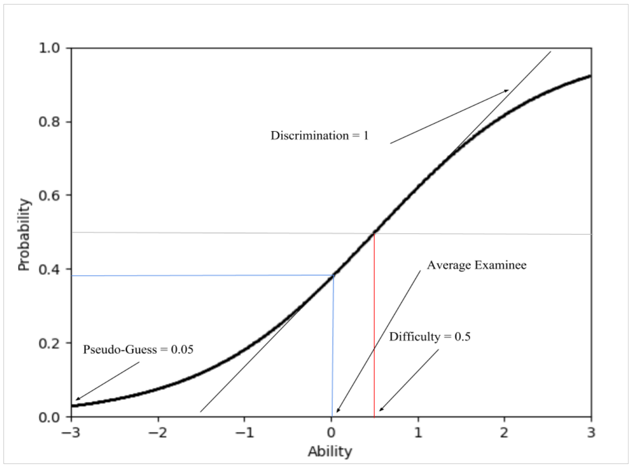

Figure 1: An example ICC

Both abilities and difficulties range from -3 to 3. An ability of -3 means the lowest ability, 0 is average, and 3 is the highest ability. As for difficulties -3 means the easiest difficulty, 0 is average, and 3 is the hardest. Using equation 1 and the difficulty of a test question, we can calculate the probability over the spectrum of all abilities ranging from -3 to 3 giving us an Item Characteristic Curve (ICC) of the question (equation 4).

$$P_i(\Theta) = \frac{e^{(\Theta_j - \beta_i)}}{1 + e^{(\Theta_j - \beta_i)}} \tag{4}$$

The ICC represents the probability of an examinee obtaining the correct response to a question. The X axis ranges across examinees' ability. The Y axis is the probability an examinee of ability(x) getting a correct response to the question, ranging from 0 to 1 where 1 means that the chance of an examinee with ability(x) getting the correct response to the given test question is 100% [7, 10]. Figure 1 shows a sample ICC of a slightly difficult question $\beta(0.5)$, where an average examinee ($\Theta = 0$) has a 40% chance of getting the answer correct.

Plotting every test question's ICC gives the distribution of test items and how difficult a test is. A good test would have a large distribution across all items. "In developing any test, our aim would be to put enough stepping-stones along the path to represent all the stepping-points useful for our testing purposes, between little development and much development" [7]. Another way to look at it is, while testing the physical strength of a group of people you will not start them off by lifting 200lbs and increasing from there. Though some people might be able to lift that at first it does not give a good representation of how weak someone might be. Someone who can lift 190lbs will be considered just as weak as someone who can lift 10lbs. The same logic applies for not having a high enough difficulty.

Another aspect when creating a valid test is looking at the fit of each question. If a question is difficult, was that difficulty because it was designed to be challenging, or was there a misunderstanding in comprehension? Calculating the likelihood ratio (LR, equation 5) [5] can help. LR is calculated by splitting up the sample into two

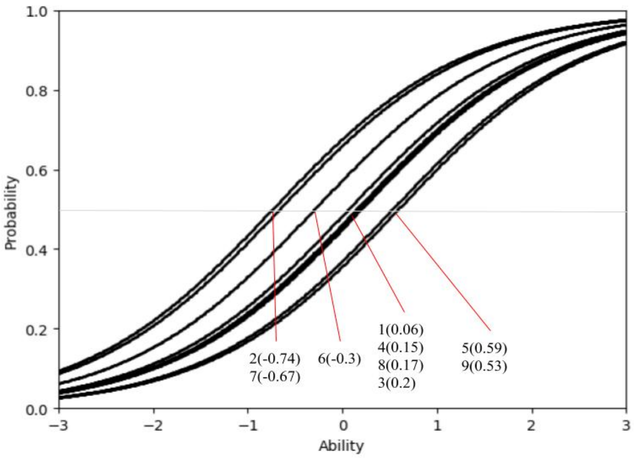

Figure 2: SCS1R ICC's

subgroups by the median score, where $\hat{\beta}_i$ consists of all students who scored below the median score and $\hat{\beta_{i_g}}$ consists of all students who scored above the median score. $l(\hat{\beta}_i)$ is the likelihood function under the overall approach. $l_g(\hat{\beta_{i_g}})$ is the likelihood function under the restricted approach. In other words, theoretically, the high ability students should have more correct responses across all test questions than low ability students. If low ability students are getting close to the same number of correct responses as high ability students to a certain question, there might be something wrong with the question. More detail can be found in [5].

$$LR = \frac{l(\hat{\beta}_i)}{\sum_{g=1}^{n-1} l_g(\hat{\beta_{i_g}})} \tag{5}$$

The Likelihood Test Statistic or z-value (Equation 6) provides a measurement as to the extent to which higher ability students perform better on a question than lower ability students. A significant z value indicates the item difficulty parameters differ across the raw score groups and that the Rasch model does not hold for that question [5]. Desirable values are routinely accepted to be those where $|z| \leq 2$ [7].

$$z = -2log(LR) = 2\sum_{g=1}^{n-1} logl_g(\hat{\beta_{i_g}}) - 2log(l(\hat{\beta}_i)) \tag{6}$$

As mentioned before, there is also a 1-par, 2-par and 3-par IRT. A 1-par IRT is commonly referred to as the Rasch model because it has the same number of parameters, but there are two significant differences. One difference is that difficulties of items in a 1-par IRT are centered around the average abilities of all examines while the difficulties of items in a Rasch model are centered around the average difficulty or all items. A second difference is that the discrimination parameter(α) of a Rasch model is set at a constant of 1 while a 1-par IRT can be set to a different constant. A 2-par IRT allows α to vary among test items. A 3-par adds in c, a lower asymptote or guessing parameter (equation 7). Changing α causes the slope of the ICC to either steepen or flatten out based on whether the question has high discrimination or not. A higher discrimination leads to a steeper slope. Changing c will cause the y asymptote to move up or down based on how easily a question can be guessed correctly [7, 10, 14].

Table 1: SCS1R Difficulties (N = 635)

Question	n+	% correct	CML
Q1	195	30.7	0.063
Q2	304	47.9	-0.740
Q3	179	28.2	0.199
Q4	185	29.1	0.148
Q5	138	21.7	0.589
Q6	241	38.0	-0.296
Q7	293	46.1	-0.665
Q8	182	28.7	0.173
Q9	144	22.7	0.528

$$P_i(\Theta) = 1 - c(\frac{e^{\alpha(\Theta_j-\beta_i)}}{1+e^{\alpha(\Theta_j-\beta_i)}}) \tag{7}$$

4 RESULTS

4.1 IRT

4.1.1 Difficulties. Table 1 shows each item and its difficulty parameter (CML) after running a Rasch Model. n+ indicates the number of students that answered the question correctly. Arranging the questions from easiest to hardest we get Q2 being the least difficult of -0.740 followed by Q7, Q6, Q1, Q8, Q3, Q9 and Q5 being the most difficult of 0.589.

All questions had less than 50% of students correctly answering the questions, labeling all questions as hard. Parker et al. [13] note that a moderate question is one where between 50% and 85% of participants answer the question correctly, while an easy question is answered correctly by more that 85% of the respondents.

4.1.2 ICC. With the difficulties calculated we are able to plot the ICC of every question. Figure 2 shows the ICC for each question. Our results indicate that there is little variation between the 9 questions. Grouping questions based on their difficulties, 4 main groups emerge. Questions 2 and 7 mark the lower end, then question 6, then questions 1, 4, 8 and 3 and questions 9 and 5 being the last and most difficult. Even though the range of difficulty is small we can still categorize students into 5 main categories; 0 questions correct, 1-2 questions correct, 3 questions correct, 4-7 questions correct, and 8+ questions correct.

4.1.3 Fit. Table 2 shows the results of running a likelihood ratio test. After splitting the population into the 2 subgroups, where the first group (indicated as low abilities in Table 2) is all students who scored below the median score of 4 and the second group (indicated as high abilities in Table 2) is those students who scored a 4 or above. An ideal test should have a significant difference between the two groups and the percentage of students who answered the question correctly for each question. Question 7 is the best fit having a z-value of 0.659. We see that 32.7% of low ability students answered

Table 2: SCS1R Frequencies and Estimates

	Total			Low Abilities			High Abilities			|z- value|
Question	n+	% correct	CML	n+	% correct	CML	n+	% correct	CML	
Q1	195	30.7	0.063	69	15.9	0.260	126	62.7	-0.190	2.229
Q2	304	47.9	-0.740	145	33.4	-0.654	159	79.1	-0.973	1.566
Q3	179	28.2	0.199	77	17.7	0.135	102	50.7	0.289	0.786
Q4	185	29.1	0.148	72	16.6	0.212	113	56.2	0.072	0.702
Q5	138	21.7	0.589	46	10.6	0.708	92	45.8	0.488	1.022
Q6	241	38.0	-0.296	114	26.3	-0.337	127	63.2	-0.210	0.666
Q7	293	46.1	-0.665	142	32.7	-0.625	151	75.1	-0.755	0.659
Q8	182	28.7	0.173	66	15.2	0.310	116	57.7	0.013	1.477
Q9	144	22.7	0.528	87	20.0	-0.008	57	28.4	1.266	6.197

that question correctly, while 75.1% of high ability students answered that question correctly. Question 9 had the worst $|z|$-value at a 6.197. We see little difference between the two groups: 20.0% of the low ability students and only 28.4% of the high ability students answering the question correctly. Since question 9's $|z|$-value is so much greater than 2 it should be either edited or removed.

4.2 Reliability

For reliability, we calculated a Cronbach's alpha of a 0.499, indicating that at least 49.9% of the total score variance was due to true score variance in this sample. With further analysis, we determined that question 9 had a negative impact on the total reliability. After removing question 9, we calculated a Cronbach's alpha of a 0.57, still below our .70 minimum cut off [12]. We note that the low Cronbach's alpha is in line with the original SCS1 reliability of 0.59 as reported by [13], suggesting a practical level of reliability in keeping with the original unmodified exam.

4.3 Comparison

We used a t-test to compare students' SCS1R scores and their final letter grades. Because we didn't plan on collecting final letter grades until later in the collection process, the total number of final letter grades that were collected was 148. Final letter grades were split up into two groups. Group 1 consisted of students received a B+ or higher(n = 102) and group 2 consisted of students who receiving a B or below(n = 46). We wanted to split students up between a B and above and a B- and below. We were unable to do so because group 2 sample size wouldn't be large enough.

Running a t-test showed that there was a significant difference of scores between students who received a B+ or above (Mean = 3.4, Median = 3, SD = 1.7) and students who received a B and below (Mean = 2.7, Median = 3, SD = 1.3) at a p-value < 0.04 which is below the minimum cut off of a 0.05 [19]. Figure 3 Shows the the box plot of both groups. While the median of both groups are the same there was a significant difference between the two means (statistic=2.1, p-value=0.039). Stating that students who received a final letter grade of a B+ or higher scored higher on the SCS1R than those who received a final letter grade of a B or below.

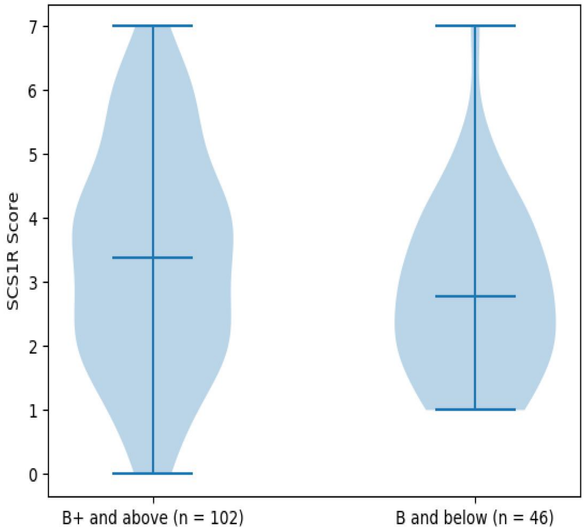

Figure 3: Violin Plot of Final letter grades and SCS1R scores (n = 148)

5 DISCUSSION

Conducting this study provided us a better understanding of the reliability of the instruments selected for use and enabled us to catch mistakes early in the data collection process for our overall study of spatial skills. Overall the SCS1R was found to be quite difficult for students, confirming the results obtained by Parker et al. [13]. Resulting data analysis suggests that this test should be further revised for difficulty and re-validated, a notion also suggested by the original authors. In our findings, beyond the difficulty level, question 9 specifically should be revised for better clarity or removed. In the case of question 1, which also fell (slightly) outside the recommended $|z|$-value, we determined that the item was necessary and beneficial to the study, while its removal caused the reliability of the SCS1R results to drop.

We where unable to reach significance while running a linear regression across all final letter grades and SCS1R scores. We chose to split students letter grades into two groups to show that the instrument is still able to differentiate between student of higher ability (course grades B+ and above) and students with lower abilities (as signified by those students earning a grade in CS1 of B and bellow). There were a few factors that play a role in why we couldn't reach significance of the chi-squared (linear regression). The biggest one is that the SCS1R is difficult. Students who received a B or below obtained scores that were not demonstrably different from guessing. There was also a strong bias when looking at final letter grades where the majority of students who participated received a grade in the A range. Another factor is that grading is not consistent across each university and their CS1 courses. Different professors have different grading scales and foci (e.g. emphasizing projects versus exams).

6 CONCLUSION

After taking a subset of the original SCS1 we were able to reduce time and redundancy, and alleviate student apprehension in taking a long and hard exam. Our Cronbach's alpha of 0.57 was close to the original Cronbach's alpha of 0.59 as reported by [13]. We also note that there was a clear correlation between student performance on the SCS1R and their performance in the CS1 course. This helped to convince us that the SCS1R exam is a reasonable predictor of student comprehension of CS1 content across the three universities data was collected.

ACKNOWLEDGEMENT

This material is based upon work partially supported by the National Science Foundation under Grant Numbers 1711780, 1711830, and 1712331. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] [n. d.]. Computer programming aptitude test. https://www.kent.ac.uk/careers/tests/computer-test.htm. ([n. d.]). Accessed: April 1, 2018.

[2] [n. d.]. CS education evaluation tools. https://csedresearch.org/tools/. ([n. d.]). Accessed: April 1, 2018.

[3] R. Alexandrowicz. 2012. "Ganz Rasch": A free software for categorical data analysis. *Social Science Computer Review* 30, 3 (2012), 369–379.

[4] E. Andersen. 1973. A goodness of fit test for the Rasch model. *Psychometrika* 38, 1 (1973), 123–140.

[5] F. Baker and S. Kim. 2004. *Item Response Theory: Parameter estimation techniques* (2nd. ed.). CRC Press.

[6] College Board. 2010. *2009 ap computer science a released exam.* The College Board, New York.

[7] T. Bond and C. Fox. 2001. *Applying The Rasch Model: Fundamental measurement in the human sciences.* Laurence Erlbaum Associates.

[8] S. Cooper, K. Wang, I. Maya, and S. Sorby. 2015. Spatial skill training in introductory computing. *ICER '15* (2015), 13–20.

[9] A. Field. 2009. *Discovering statistics using SPSS.* Sage publications.

[10] D. Harris. 1989. Comparison of 1-, 2- and 3-Parameter IRT Models. *Educational Measurement: Issues and Practice* 8, 1 (1989), 35–41.

[11] L. Lacher, A. Jiang, Y. Zhang, and M. Lewis. 2017. Aptitude and previous experience in cs1 classes. *Int'l Conf. Frontiers in Education: CS and CE* 17 (2017), 87–91.

[12] C. Lance, M. Butts, and L. Michels. 2006. The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods* 9, 2 (2006), 202–220.

[13] M. Parker, M. Guzdial, and S. Engleman. 2016. Replication, validation, and use of a language independent cs1 knowledge assessment. *ICER '16* (2016), 93–101.

[14] Ayala R.J. de. 2009. *The theory and practice of item response theory.* The Guilford Press.

[15] A. Sahin and A. Duygu. 2016. The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory Practice* 17, 1 (2016), 321–335.

[16] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (1908), 1–25.

[17] A. Tew, B. Dorn, and O. Schneider. 2012. Toward a validated computing attitudes survey. *ICER '12* (2012), 135–142.

[18] A. Tew and M. Guzdial. 2011. The FCS1: A language independent assessment of cs1 knowledge. *SIGCSE '11* (2011), 111–116.

[19] R. Wasserstein and N. Lazar. 2016. The ASA's Statement on p-Values. *The American Statistician* 70, 2 (2016), 129–133. https://doi.org/10.1080/00031305.2016.1154108

[20] S. Y. Yoon. 2011. Revised Purdue Spatial Visualization Test: Visualization of rotations (Revised PSVT:R) [Psychometric Instrument]. (2011).