Validating a Language-Independent CS1 Learning Outcomes Assessment

Aditya Jain
The University of Nebraska - Lincoln
Lincoln, Nebraska, USA
ajain9@unl.edu

Ryan Bockmon
The Roux Institute at Northeastern University
Portland, Maine, USA
r.bockmon@northeastern.edu

Chris Bourke
The University of Nebraska - Lincoln
Lincoln, Nebraska, USA
chris.bourke@unl.edu

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

ABSTRACT

Assessing learning outcomes in computer science education is essential as it is an indicator of student progress, the effectiveness of teaching methods, and areas for improvement. Aptitude tests have been widely used to measure these learning outcomes; however, they are not without their issues with reliability, difficulty, and applicability across courses and institutions. To address these issues, this study aims to contribute to the development of a reliable, language-independent testing instrument that accurately evaluates students' performance, capabilities, and grasp of the learning outcomes from an introductory computer science course. In this study, we employed the Second Computer Science 1 Exam Revised version 2 (SCS1Rv2) as a post-assessment tool to measure learning outcomes. The SCS1Rv2 was administered in three CS1 course sections, and the results were compared with the final grades of the students. The validation of the SCS1Rv2 was done using Item Response Theory where the test was assessed for its difficulty and reliability. We found that the SCS1Rv2 is a reasonable predictor of course learning outcomes. The intent of this study is to aid in the creation of a standardized, reliable, and effective testing instrument that can be used across different courses and institutions. The SCS1Rv2 has the potential to be a valuable tool in its development.

CCS CONCEPTS

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

KEYWORDS

CS1, Assessment, Validation, Item Response Theory

ACM Reference Format:
Aditya Jain, Ryan Bockmon, Chris Bourke, and Stephen Cooper. 2023. Validating a Language-Independent CS1 Learning Outcomes Assessment. In *Proceedings of the ACM Conference on Global Computing Education Vol 1 (CompEd 2023), December 5–9, 2023, Hyderabad, India.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3576882.3617910

CompEd 2023, December 5–9, 2023, Hyderabad, India

ACM ISBN 979-8-4007-0048-4/23/12…$15.00
https://doi.org/10.1145/3576882.3617910

1 INTRODUCTION

In computer science education, testing instruments are used to assess students' learning and progress. Measuring learning outcomes is important as it is not only an indicator of student progress but also allows the instructor to evaluate their teaching methods and identify areas of improvement in their course. In introductory computer science courses, measuring learning outcomes can be challenging due to the nature of these introductory courses. Typically, a wide range of material and topics are covered in CS1 courses as they serve as a comprehensive introduction to the field for first-time students. There is a need for a well-designed, standardized, valid instrument that can reliably assess learning outcomes for an introductory computer science course.

Computer science aptitude tests are commonly employed to gauge students' skill sets and ability to solve problems and complete different tasks related to the course material. A well-designed introductory computer science test should dependably measure core concepts such as numerical and logical reasoning, algorithm comprehension, and program simulation, which represent overall aptitude for the subject. However, the large variety of CS1 courses taught, including different programming language focuses, creates an obstacle to reliably measuring students' aptitude across courses or institutions. A standardized, language-independent test would be able to address this issue. Additionally, an assessment like so would help ensure that the current curriculum in, or even across, universities is adequately achieving desired learning outcomes.

The testing instrument used in this study was the Second Computer Science 1 Exam Revised version 2 (SCS1Rv2). The SCS1Rv2 was used and validated previously, by Bockmon et al. [3] as a pre-assessment in CS1 courses along with the Computational Thinking Concepts and Skills Test. The combined test was called the Placement Skill Inventory (PSIv1) and was used to help students decide to enroll in a CS0 or a CS1 course. This study aims to validate and evaluate the SCS1Rv2 as a post-assessment. Our goal is to review if this instrument serves to measure learning outcomes for a CS1 course. This study is conducted in an effort to contribute to the development of valid, reliable, accurate computer science testing instruments that can be used in pedagogical settings across CS1 courses, regardless of language or institution.

2 RELATED WORK

Over the past several years there has been a great effort to validate a language-independent introductory computing programming test. Starting in 2011, Tew and Guzdial created the Foundational CS1 Assessment instrument (FCS1) [9] which was an attempt to create a CS assessment test independent of any single programming language. The FCS1 tests student ability using pseudocode written in the style of imperative languages. The validation process was conducted through a multi-step process that included an expert review panel, and a large-scale comparison between the FCS1 and language-dependent isomorphic tests [6, 9]. 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 has the highest correlations with a Pearson's $r = 0.511$, $p < 0.001$. Python originally had the lowest correlation, but after splitting into two subgroups results showed that the CS-python focused section had a Pearson's $r = 0.679$, $p < 0.001$ and Media-Python had a Pearson's $r = 0.262$, $p < 0.001$.

The Second CS1 exam (SCS1) [7], a successor of the FCS1, was created in 2016 and consisted of 27 multiple-choice questions. 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 of $r = 0.566$, $p < 0.001$. 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 0.65 with their FCS1 having a Cronbach's alpha of 0.53 and the SCS1 with a Cronbach's alpha of 0.59 [7]. 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 its length, as it took students a long time (between 2 - 3 hours) to complete it.

Bockmon et al. [5] revised the SCS1 down to a total of nine multiple choice questions as they needed a test that was language-independent and shorter to complete. The nine questions that they chose to keep reflected several major topics likely to be covered in an introductory computing course. These nine questions were then called the Second CS1 exam - Revised (SCS1R).

Bockmon et al. [3] then went on to combine the SCS1R with the Computational Thinking Concepts and Skills Test [8] to create the Placement Skill Inventory (PSIv1) a CS0/CS1 placement exam. This experiment was administered as a pre-assessment before the semester began and was designed to help students decide whether to enroll in either a CS0/CS1 course. The PSIv1 consisted of 24 computational thinking and CS programming comprehension multiple-choice questions.

3 METHODS

Data was collected across three different introductory computing courses at a large midwest R1 university. Two different programming languages (Java or Python) were used across the three sections. Participation was mandatory for all students who were enrolled in one of those sections, in an effort to avoid participation bias [4]. There was a total of 219 students who participated across all sections. The SCS1Rv2 was administered at the end of the semester and was part of students' final grade in the course. Depending on which section they were enrolled in, students took the SCS1Rv2 as either part of their final exam or their last homework assignment for the course. All students were graded based on how well they did on the test itself to increase motivation and incentivize students to give a true effort.

Students who took the SCS1Rv2 as part of their final were allowed to take it as an open book/open note test. Students had a 24-hour open window in which to complete it. No collaboration was allowed on either part.

4 SCS1RV2 DESIGN

The original SCS1R had nine programming language independent multiple choice questions that covered basic topics that all introductory computing courses would likely cover, such as fundamentals, logical operators, conditionals, definite and indefinite loops, arrays, function parameters, function return values, recursion, and object-oriented basics. It was designed to take students around 30-45 minutes to complete, to mitigate any effects of fatigue on examinees. During the original validation of the SCS1R, there was one question that was identified to be a "poor fitting" question and was noted to be revised or removed. That question is not included in the SCS1Rv2. The authors also noted that the SCS1R was a difficult test for students to score well on, with the average score being just above guessing. To account for the difficulty of the original test, four questions were taken from the CS-AP exam [2] that were thought to be easier questions. These four questions were then converted to pseudocode to match the language-independent format. These questions were added to the remaining eight SCS1R questions (removing the question with poor fit) to create the SCS1Rv2. The SCS1Rv2 has 12 multiple-choice questions and is still designed to take students 30-45 minutes to complete.

5 ANALYSIS

The validation of SCS1Rv2 was conducted using Item Response Theory (IRT) [1]. Using IRT, we can understand the difficulty and fit for each item/question. The difficulty of a question is determined by the number of examinees that correctly answered the question. The conditional maximum likelihood (CML) estimation was used to measure difficulties, and it ranges on a scale of -3 to 3. A question with a computed difficulty/CML of -3 had a very low difficulty for the examinees, while a difficulty of 0 represented an average level of difficulty, and a 3 denoted a very high difficulty level. A well-designed test should include questions that span different levels of difficulty, allowing for the assessment of the examinee's skills across a broad spectrum.

The fit of a question was measured by the likelihood test statistic or $|z|$-value. This statistic was determined by how well the particular question can differentiate low ability examinees from high ability examinees. A significant $|z|$-value implies that IRT does not apply to the question since the question's difficulty parameters vary between the raw score groups. For IRT to hold, a desired $|z|$-value is where $|z| \leq 2$.

IRT takes into consideration the examinee's ability and calculates this value determined by the examinee's performance on the entire test. Similar to a question's difficulty, an examinee's ability ranges from -3 to 3. An ability of -3 represents very low ability, 0 means average ability, and 3 corresponds with very high ability.

With the question's difficulty (β) and fit, IRT aims to mathematically represent the relationship between an examinee's ability (Θ), and the probability (P) of the examinee answering the question (i) correctly, seen below.

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

Using this formula, we can generate an Item Characteristic Curve (ICC) for each question. The ICC models the probability of an examinee correctly answering a question. The examinee's ability is represented on the X-axis. The probability that an examinee with an ability (x) will correctly answer the question is represented on the Y-axis.

To dive into the participants' SCS1Rv2 scores, measures of central tendency were computed first. Tests for normality were run to determine if participants' scores and final grades were normally distributed. A t-test for independence was used to test for a significant difference between SCS1Rv2 scores and final grades for the participants, both of which were converted into percentages. We additionally plotted for correlation between the SCS1Rv2 scores and final grades. We then split all the participants into two groups, aiming to match the Low and High-Ability split done before in IRT. For these two groups, measures of central tendency were calculated, t-tests were utilized to compare the SCS1Rv2 scores of both groups, and the correlation between SCS1Rv2 scores and final grades was determined. Throughout this study, an alpha of 0.01 is used as the cutoff for significance, and a p-value of less than 0.01 indicates statistical significance in the results.

6 RESULTS

6.1 Item Response Theory

6.1.1 Difficulties. Seen in Table 1 are each of the 12 questions in SCS1Rv2 and their computed metrics by running IRT. The results for all participants are shown in the section of the table labeled "Total". The sections marked "Low Abilities" and "High Abilities" represent the portions of the participants in each of those respective groups, classified by their score on the entire SCS1Rv2. The final column lists the calculated $|z|$-value for the question, a measure of its fit. Within each section, n+ indicates the number of students that correctly responded to the question, and this value represented as a percentage is also included. Additionally, CML denotes the question's difficulty for the section.

Examining the difficulties of each question, it is clear that Q8 is the most difficult, with the lowest percentage of correct responses across all groups and a CML of 1.073. Q3 was a close second with a CML of 1.024. The easiest questions were Q10 and Q11, both having a CML of -1.442. Also, Q10 and Q11 had the highest percentage of correct responses, by a considerable amount, among all participants. In order from least difficult to most difficult, we have Q10, Q11, Q2, Q9, Q12, Q7, Q4, Q6, Q1, Q5, Q3, Q8. The range of the percentages of correct answers in the Low Abilities group was much larger than that of the High Abilities group. In the Low Abilities group, this percentage ranged from [34.6%, 87.9%], with a mean of 57.8%. The range in the High Abilities group was [82.1%, 97.3%], with a mean of 91.8%.

6.1.2 Item Characteristic Curve. The ICC displays the probabilities of a participant with a given certain ability to correctly answer the 12 questions in SCS1Rv2, plotted in Figure 1. Each question's ICC is labeled, and the order of these curves follows the order of questions, ranking them in order of increasing difficulty. Although some questions have a similar level of difficulty as others, there is no one distinct level of difficulty for this exam. This illustrates that the 12 questions from SCS1Rv2 have a widespread range of difficulties, with some being easier (Q10, Q11, Q9, Q2), others being average difficulty (Q12, Q7, Q4), and others being more challenging (Q1, Q5, Q3, Q8).

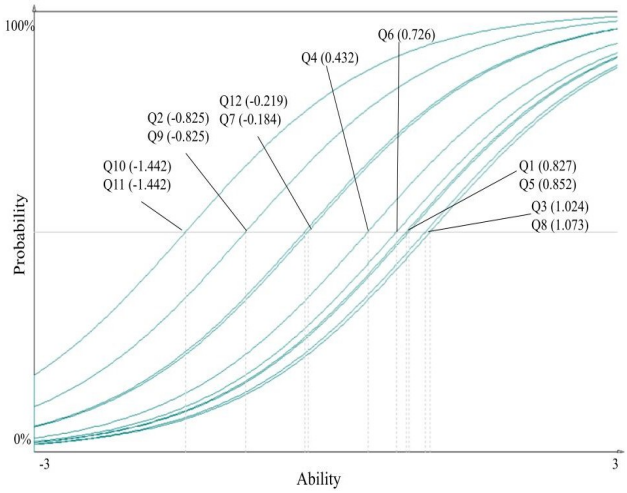

Figure 1: SCS1Rv2 Item Characteristic Curve for 12 SCS1Rv2 Questions

6.1.3 Fit. The $|z|$-value column measures each of the 12 questions' fit. In other words, it is the measure of how well each question can distinguish a participant from the Low Abilities group and the High Abilities group. Looking at these results from this column, only Q10 has a $|z|$-value ≥ 2, deeming it with poor fit. A closer examination of Q10 reveals that 87.9% of the Low Ability participants answered correctly and 95.5% of the High Ability participants correctly responded. The difference in these two percentages is not substantial enough to distinguish the two groups. In fact, Q10 has the smallest difference between the two groups in the percentage of correct responses for a particular question. Given all this, Q10 should either be revised or removed from SCS1Rv2.

Conversely, the question with the best fit is Q7, with a $|z|$-value = 0.345. This question had a large enough variation among the Low and High Abilities groups, 63.6% and 94.6%, respectively, to be able to differentiate between them. Moreover, there were several other questions with a $|z|$-value similar to Q7: Q2 (0.379), Q9 (0.379), Q8 (0.383), Q1 (0.384), and Q3 (0.428). Having very a low $|z|$-value gives

Table 1: Item Response Theory Results for SCS1Rv2

Question	Total (n = 219)			Low Abilities (n = 107)			High Abilities (n = 112)			$\lvert z \rvert$-value
	n+	%Correct	CML	n+	%Correct	CML	n+	%Correct	CML	
Q1	139	63.5	0.827	43	40.2	0.858	96	85.7	0.723	0.384
Q2	190	86.8	-0.825	81	75.7	-0.794	109	97.3	-1.026	0.379
Q3	131	59.8	1.024	38	35.5	1.062	93	83.0	0.915	0.428
Q4	154	70.3	0.432	49	45.8	0.621	105	93.8	-0.157	1.769
Q5	138	63.0	0.852	41	38.3	0.939	97	86.6	0.652	0.800
Q6	143	65.3	0.726	45	42.1	0.779	98	87.5	0.557	0.557
Q7	174	79.5	-0.184	68	63.6	-0.156	106	94.6	-0.317	0.345
Q8	129	58.9	1.073	37	34.6	1.103	92	82.1	0.973	0.383
Q9	190	86.8	-0.825	81	75.7	-0.794	109	97.3	-1.026	0.379
Q10	201	91.8	-1.442	94	87.9	-1.705	107	95.5	-0.505	2.227
Q11	201	91.8	-1.442	92	86.0	-1.530	109	97.3	-1.026	0.798
Q12	175	79.9	-0.219	73	68.2	-0.384	102	91.1	0.216	1.486

all of these questions a strong fit, as they are capable of correctly classifying participants into the Low Abilities or High Abilities group.

6.2 Grade Distribution

In this study, the SCS1Rv2 had a mean score of 74.8% (about 9/12 questions), a median score of 83.3% (about 10/12), and a standard deviation of 20.7%. After verifying the scores and grades were normally distributed, running a t-test between the participants' SCS1Rv2 scores and their final grades resulted in a t-statistic of -5.78 and a p-value < 0.01. This shows a statistically significant difference between the SCS1Rv2 scores and final grades of all the participants in this study. However, there was a moderately strong correlation between these two variables, with a Pearson's $r = 0.47$, $p < 0.01$. Figure 2 shows the scatter plot for these two variables, including the mean score at each distinct SCS1Rv2 score (marked with the red triangle), along with the best-fit linear regression line.

All participants were split into two groups based on their final grade, in an effort to match the two different ability groups split when running IRT. Participants with a final grade at or above 88%, corresponding with a grade greater than or equal to B+, were put into Group 1. Those with a final grade below 88%, corresponding with a grade less than a B+, made up Group 2. Group 1 had 109 participants; Group 2 had 110 participants. Group 1 had a mean of 83.4%, median of 91.7%, and standard deviation of 15.6%. Group 2 had a mean of 66.2%, median of 66.7%, and standard deviation of 21.6%. From this alone, it is clear that Group 1, participants with a final grade above a B+, scored much higher on the SCS1Rv2, getting around 2-3 more questions correct than their counterparts in Group 2. The correlation coefficient representing the relationship between final grades and SCS1Rv2 scores was $r = 0.37$ for Group 1. The correlation between these two variables was not as strong in Group 2, coming out to $r = 0.26$. Running a t-test between the final grades and SCS1Rv2 scores in each group showed that there was a statistically significant difference for both, having a p-value < 0.01.

To compare the SCS1Rv2 scores for both groups, a t-test was run that resulted in a p-value < 0.01. This again backs up the claim that the participants in Group 1 significantly outperformed Group 2 on

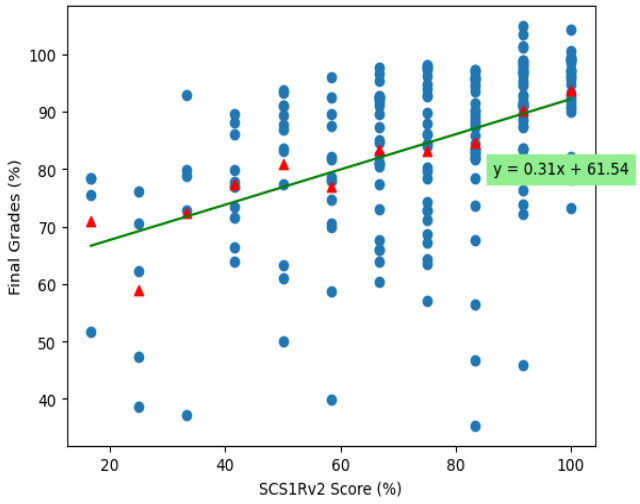

Figure 2: Participants' Final Grade vs SCS1Rv2 Score Scatter Plot

the SCS1Rv2. The difference in the standard deviations between the group is noteworthy, with Group 2's being greater, hinting at a larger spread of scores. This can be seen in Figure 3, which plots the distribution of SCS1Rv2 scores for both groups side-by-side as violin plots. It is clear that the median for Group 1 is substantially higher than that of Group 2, but also the scores for Group 1 are much more concentrated than Group 2. Group 1's scores span from [33.3%-100%], while Group 2's scores vary from [16.7%-100%], and this is supported by the comparatively larger standard deviation for Group 2.

7 DISCUSSION

By completing this study, we have gained valuable insight into the dependability and validity of the SCS1Rv2. Looking at the average difficulty of all the questions in the assessment (mean CML = -0.00025), a mean examinee score of 74, 8%, and a median score

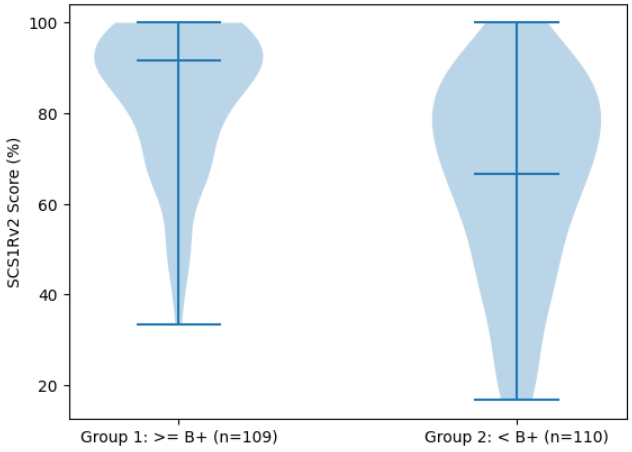

Figure 3: Violin Plots of Participants' SCS1Rv2 Scores split on B+ final grade

of 83.3%, we can infer that the SCS1Rv2 was an exam of relatively average difficulty for the examinees, possibly leaning towards being slightly easier. Moreover, visible from the ICC, Figure 1, the wide spread of difficulties of questions, from [−1.442-1.073], contributed to the strength of this assessment. None of the questions were too easy, or too difficult, yet there was enough variation in their difficulties that the assessment could distinguish between low ability examinees and high ability examinees. The significant difference between the means of the High Abilities group/Group 1, and the Low Abilities group/Group 2, is supported and reflected by 11 out of 12 questions having a good fit ($|z|$-value < 2). In other words, all but one question performed well to differentiate examinees as having a high or low ability. Coupled with a moderately strong correlation between SCS1Rv2 scores and final grades (Figure 2) we can reasonably deduce that the SCS1Rv2 performed well at gauging learning outcomes for the CS1 course of this study.

Focusing on the validation of each question individually from the SCS1Rv2, gave us a deeper understanding of the overall assessment. Q10 was tied for the easiest question but also was the only one with a poor fit to IRT, and therefore, should be revised or removed from the SCS1Rv2. Because it is tied with Q11 for the easiest question, the complete removal of Q10 is unlikely to significantly change the overall difficulty of this assessment.

Four questions were added to the original SCS1R [5] to create the SCS1Rv2 [3] (Q9 - Q12, in this study). These four questions were included in an effort to make the assessment easier and better at differentiating students as having High or Low abilities. The results of this study corroborate this, as those four questions (Q9 - Q12) were indeed easier for the examinees. The data in Table 1 demonstrates that in the Low Abilities group, the average percentage of examinees getting correct responses when only looking at the first eight questions is 49.0%. However, the average percentage of examinees in this group getting the last four questions correct is 79.5%, which bumps up the average percentage of examinees responding correctly to all 12 questions to 57, 8%. The same trend is reflected in the High Abilities group as well, with the average percentage of examinees in the group getting the correct response for the first eight questions being 90.1%, the last four questions being 95.3%, and all 12 questions being 91.8%. Thus, examinees find the last four questions in the SCS1Rv2 relatively easy, which contributes to the overall wide distribution of difficulties across the assessment.

8 LIMITATIONS

Despite the significant findings, this study has certain limitations that should be acknowledged and considered in order to interpret the findings accurately. This study was run entirely at one university. While we obtained a sizeable sample, there is a possibility that a more diverse sample from various universities might result in different outcomes. The design of this study was with the intent of reducing participation bias within our experiment sample. Having the SCS1Rv2 given as an attachment to the examinees' final grade was an effort to invoke self-motivation among students to perform well. However, since the study was conducted at the end of the semester, students may not give full effort towards this assessment if they are content with their current course grade and can afford a lower score on the final exam, homework, or however the SCS1Rv2 was administered. Also, by administering the SCS1Rv2 at the end of the semester, it fails to capture any students who had dropped out of the course anywhere in the middle of the semester. Additionally, while the students were graded based on their performance on the test, there was no pressing time restriction, as they had a 24-hour window to complete it.

Related to the structure of this study, the SCS1Rv2 was validated before, by Bockmon et al. [3], as a pre-assessment to a CS1 course. This study validates the assessment as a post-assessment but on a different sample of students; thus there is no validation of the SCS1Rv2 in the pre/post format. Lastly, while the SCS1Rv2 was written to be programming-language independent, it was written in English. Any examinee for who English was not their primary language may have faced additional difficulties in completing the assessment.

9 CONCLUSION

The development and validation of research instruments, like assessments, is a continuous and dynamic process with the aim to make these instruments stronger and more reliable. This study provides a deeper understanding of the SCS1Rv2 as an instrument to measure learning outcomes for a CS1 course. The majority of this assessment effectively differentiated performance levels among the examinees. Considering the strong correlation between SCS1Rv2 scores and final grades, well-fit questions, and a wide range of question difficulties, we can conclude the testing instrument is a suitable and credible indicator for assessing learning outcomes by the end of a CS1 course. We strive to build upon the SCS1Rv2 to make it a more robust and reliable instrument for future applications in introductory CS courses.

ACKNOWLEDGMENTS

We would like to thank Dr. Witawas Srisa-An, for helping us collect data from his course.

REFERENCES

[1] Frank Baker and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques* (2 ed.). CRC Press.

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

[3] Ryan Bockmon and Chris Bourke. 2023. Validation of the Placement Skill Inventory: A CS0/CS1 Placement Exam. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 39–45. https://doi.org/10.1145/3545945.3569738

[4] Ryan Bockmon and Stephen Cooper. 2022. What's Your Placebo? *Commun. ACM* 65, 10 (sep 2022), 31–33. https://doi.org/10.1145/3528085

[5] 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* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 552–557. https://doi.org/10.1145/3287324.3287372

[6] Allison Elliott Tew, Brian Dorn, and Oliver Schneider. 2012. Toward a Validated Computing Attitudes Survey. In *Proceedings of the Ninth Annual International Conference on International Computing Education Research* (Auckland, New Zealand) *(ICER '12)*. Association for Computing Machinery, New York, NY, USA, 135–142. https://doi.org/10.1145/2361276.2361303

[7] Miranda C. Parker, Mark Guzdial, and Shelly Engleman. 2016. Replication, Validation, and Use of a Language Independent CS1 Knowledge Assessment. In *Proceedings of the 2016 ACM Conference on International Computing Education Research* (Melbourne, VIC, Australia) *(ICER '16)*. Association for Computing Machinery, New York, NY, USA, 93–101. https://doi.org/10.1145/2960310.2960316

[8] Markeya S Peteranetz and Anthony D Albano. 2020. Development and Evaluation of the Nebraska Assessment of Computing Knowledge. *Frontiers in Computer Science* 2 (2020), 11.

[9] Allison Elliott Tew and Mark Guzdial. 2011. The FCS1: A Language Independent Assessment of CS1 Knowledge *(SIGCSE '11)*. Association for Computing Machinery, New York, NY, USA, 111–116. https://doi.org/10.1145/1953163.1953200