An Item Response Theory Evaluation of a Language-Independent CS1 Knowledge Assessment

Benjamin Xie
University of Washington
The Information School, DUB Group
Seattle, Washington, USA
bxie@uw.edu

Matthew J. Davidson
University of Washington
College of Education
Seattle, Washington, USA
mattjd@uw.edu

Min Li
University of Washington
College of Education
Seattle, Washington, USA
minli@uw.edu

Amy J. Ko
University of Washington
The Information School, DUB Group
Seattle, Washington, USA
ajko@uw.edu

ABSTRACT

Tests serve an important role in computing education, measuring achievement and differentiating between learners with varying knowledge. But tests may have flaws that confuse learners or may be too difficult or easy, making test scores less valid and reliable. We analyzed the Second Computer Science 1 (SCS1) concept inventory, a widely used assessment of introductory computer science (CS1) knowledge, for such flaws. The prior validation study of the SCS1 used Classical Test Theory and was unable to determine whether differences in scores were a result of question properties or learner knowledge. We extended this validation by modeling question difficulty and learner knowledge separately with Item Response Theory (IRT) and performing expert review on problematic questions. We found that three questions measured knowledge that was unrelated to the rest of the SCS1, and four questions were too difficult for our sample of 489 undergrads from two universities.

CCS CONCEPTS

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

KEYWORDS

validity; assessment; item response theory; CS1; concept inventory

ACM Reference Format:
Benjamin Xie, Matthew J. Davidson, Min Li, and Amy J. Ko. 2019. An Item Response Theory Evaluation of a Language-Independent CS1 Knowledge Assessment. 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, 7 pages. https://doi.org/10.1145/3287324.3287370

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

ACM ISBN 978-1-4503-5890-3/19/02…$15.00
https://doi.org/10.1145/3287324.3287370

1 INTRODUCTION: IMPROVING MEASUREMENT OF CS KNOWLEDGE

Assessment is key to measuring knowledge, skills, and abilities related to computing [34]. For an instructor, assessments can differentiate between students' level of knowledge, helping personalize instruction. For learners, assessments can help them identify what they do and do not understand. For researchers and instructional designers, assessments can help measure the effect of interventions.

One recent and widely-used assessment is the Second Computer Science 1 (*SCS1*) concept inventory. The SCS1 was developed by Parker, Guzdial, & Engleman [26] as a replication to another concept inventory [33, 35]. The SCS1 is a multiple-choice exam which covers introductory computer science (*CS1*) knowledge, but does so with a language-independent pseudo-code. In the three years since the SCS1 was published, researchers have used it to measure students' CS1 knowledge before and after a course [12, 26, 36, 37] as part of pre-test/post-test studies [4], CS teachers' knowledge [26], and pre-CS1 students' knowledge [23].

The SCS1 is one of few CS concept inventories that has undergone validation to evaluate the validity and reliability of its test scores [26]. *Validation* is a process that involves evaluating the plausibility and appropriateness of proposed interpretations and uses of assessment scores [11, 13]. The validation process involves *iteratively* developing an assessment and an argument that specifies how to interpret assessment scores, and then challenging the argument to identify further improvements to the argument. Validation studies can identify if a test is too easy or too difficult for the target population [1], as well as whether surface features of questions [6] (e.g. wording of problem, style of code) confound the score [5, 16].

Through validation, we can produce evidence (empirical, theoretical, and argumentative) to support the interpretations of test scores for a proposed use. This evidence can better inform us as to which test questions are most effective at differentiating high- and low-performing students, helping us to further refine both specific questions and the set of questions included in the test as a whole. Ultimately, this evidence builds a case for the validity of interpretations of SCS1 scores as a measure of CS1 knowledge.

Validation is an iterative process, so we extended the prior validation study of the SCS1 to better understand how it measures

test-takers. Parker, Guzdial, & Engleman [26] evaluated the SCS1 following Classical Test Theory (CTT)[1] [1], but CTT has limitations which make results challenging to interpret and generalize.

In this paper, we extended the validation of the SCS1 using Item Response Theory (IRT) [8], a widely used technique in the field of psychometrics. By using IRT, we can produce evidence on the quality of the SCS1 which can better generalize to the population of CS1 learners and distinguish between question difficulty and learner knowledge. We sought to answer the following questions: **RQ1**: *Do all the SCS1 questions measure the same underlying construct (CS1 knowledge)?* **RQ2**: *For what levels of CS1 knowledge does the SCS1 measure well?* **RQ3**: *How closely do the difficulty levels of the SCS1 questions align with the knowledge levels of our sample group?* **RQ4**: *What do the response patterns of problematic questions reveal?*

2 IRT VS. CLASSICAL TEST THEORY

Classical test theory (CTT) assumes a model where a learner's observed score is a combination of their unobserved *true score* and a certain amount of measurement error [24]. CTT attempts to measure this error by deriving test parameters on the basis of *total* scores for the learners in the sample [1]. CTT is useful as an initial analysis of data [11] but has three major limitations: 1) statistics such as test reliability, item difficulty, and item discrimination are *sample-dependent* [1, 38], 2) a learner's test score is an unspecified (at least in CTT) combination of a learner's knowledge and question properties (e.g. difficulty) [38]; 3) the model of observed score being a composite of true score and error is *unfalsifiable* [20].

In contrast, Item Response Theory [8, 20] provides more sample-agnostic statistics which estimate learner knowledge and question properties separately using falsifiable models. IRT analyzes each question and each learners' performance on a question separately, estimating question-level and test-level parameters, as well as learners' knowledge levels. This provides unique estimates of the difficulty and discrimination of each question for learners of different knowledge levels. Furthermore, these estimates generalize beyond the specific sample of learners. It does this by estimating the correspondence between unobserved latent variables (e.g. people's CS1 knowledge, difficulty of questions) and observable evidence of knowledge (e.g., people's responses to questions). By fitting response data to a model (e.g. logistic or multinomial), we can estimate question parameters (e.g. difficulty) with fewer assumptions about the characteristics of the sample. We can also make predictive statements about learner performance based on knowledge level.

IRT helps reason about the relationship between question difficulty and learner knowledge by placing people and questions on the *same* (typically unidimensional) continuum. It is centered at 0, which represents the ability level (herein *knowledge level*[2]) for the average test-taker of the population. If a test-taker's predicted knowledge level is greater than the difficulty of the question, they are more likely to get the question correct. Figure 1 shows an example of such a continuum. This continuum helps model the relationship between a learner's latent knowledge level and their observed item performance as a monotonically increasing function.

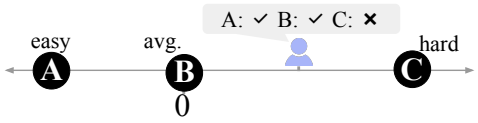

Figure 1: Representation of latent variable continuum with 3 questions (*A, B, C*). Because this learner's knowledge level is greater than the difficulties of *A* and *B*, we would predict they get those questions correct. Their level is lower than *C*'s difficulty, so we predict they get that question wrong. Zero reflects the knowledge level for the average test-taker.

Within IRT, a useful test question differentiates well between people located at different points on the continuum (having different knowledge levels) [8]. A question differentiates best between learners who have knowledge levels nearer to its difficulty. For most IRT models, the estimated difficulty of a question is the knowledge level at which a learner has an even chance (i.e. 50% probability) of answering correctly. Referring back to Figure 1, question *B* would better differentiate between learners with average knowledge levels (around 0), meaning some people would get B correct and others would not. In contrast, question *A* would be too easy to provide useful information about learners with an average knowledge level (almost everyone would get *A* correct). In addition to difficulty, some IRT models (including three models fitted below) estimate discrimination for each test question. *Discrimination* is how well a question distinguishes between knowledge levels–a high discrimination value means the probability of a learner answering correctly changes significantly based on their knowledge level. With discrimination, we can see if questions at the same difficulty level provide more or less information about learners' knowledge.

3 DATA: HUNDREDS OF SCS1 RESPONSES

The data we analyzed with IRT were responses to the SCS1, a 27 question multiple choice assessment of CS1 knowledge [26, 35]. The SCS1 covered nine concepts: *basics*, *conditionals*, *definite/for loops*, *indefinite/while loops*, *logical operators*, *arrays*, *recursion*, *function parameters*, and *function return values*. We began with SCS1 responses from 507 undergrad students about to begin 1 of 3 different courses at the University of Washington, Seattle (UW) and Georgia Institute of Technology (GT), large public universities in the United States. The data were collected for studies which used the SCS1 as a pre- and post-test to measure learning outcomes for different CS courses. We choose to analyze only pre-test data so the differences in interventions would not affect responses. Two of these courses (at different universities) administered the SCS1 within the first week of a CS1-level course, so we referred to these learners as *pre-CS1*. The third course administered the SCS1 within the first week of a CS2-level course (client-side web development), so we referred to these students as *pre-CS2*.

We only considered responses from test-takers who spent 10-70 minutes taking the SCS1. We chose the lower bound in the time limit to attempt to filter out random guessing; we choose the upper bound because the SCS1 is only supposed to take one hour.

[1] Sec. 3.2 of [26] mentioned "IRT analysis" of the SCS1, but we interpreted this as item-level analysis that follows CTT.

[2] We avoided using common IRT terms for comprehensibility. We will refer to items as *questions*, the instrument by name or as a *test*, the latent construct as *CS1 knowledge*, the item parameter as *difficulty* and the person parameter as *knowledge level*, or *knowledge/difficulty* when showing the latent variable continuum.

We also only considered test-takers who attempted at least 10 questions. The online format of the SCS1 made it such that learners could only see one question at a time and could not continue until they selected an answer, potentially forcing them to guess. So as soon as a test-taker stopped responding to a question, all further questions were left unanswered. Thus, the number of non-responses was monotonically increasing. So, we used the elbow method [29] to identify a cutoff of 10; after filtering by time and attempted questions, the number of responses decreased from 507 to 489.

Our sample had 118 responses from UW (54 pre-CS1, 64 pre-CS2) and 371 responses from GT (all pre-CSE1). While most of the sample were about to begin a CS1 course, surveys indicated that many had prior programming experience. At least half of the pre-CS1 learners from UW had prior programming experience (27/54 respondents), typically in the form of a high school course or independent learning. We only had demographic information for learners from UW, which was self-reported and had some non-response. Half (60/118) of the UW students identified as male, 38% as female, and 1 student as non-binary; the ratio of male to female was approximately equal between the pre-CS1 and pre-CS2 courses. Over 70% (87/118) were between 18-22 years old, with the remainder being older; most of the older respondents were pre-CS2. Survey respondents were predominantly Asian/Pacific Islander 46% (55/118) or 31% White/Caucasian, with only 6% reporting as minority or multi-ethnic. Java was most commonly known programming language, with 55% (66/118) of respondents reporting at least minimal proficiency in it; a majority of them were pre-CS2. We had no demographic data from GT.

4 IRT ANALYSIS: MODELING W/ 2PL

To understand how well the SCS1 measured achievement, we verified IRT assumptions then fit the response data to a logistic model.

4.1 Verifying IRT assumptions

4.1.1 Verifying conditional item independence. This assumption is that responses to a question are independent of responses to any other item, conditional on the learner's achievement level. This assumption enables us to calculate the probability of a response pattern by taking the product of the product of the probabilities of each individual response.

We justify conditional item independence through argumentation. While carryover effects [1] between questions are to some extent inevitable, we argue that conditional item independence is still valid because the online medium of the SCS1 exam only allowed respondents to see one question at a time. Furthermore, navigating back to previously attempted questions was difficult, as respondents would have had to go linearly back across previous questions and the questions pertaining to the same concept tended to be separate from each other. So, we assumed conditional item independence to be true.

4.1.2 RQ1: Verifying unidimensionality with CFA. The second assumption is *unidimensionality*, which states that only one unobserved "trait" is being measured by all of the questions [8, 31]. While the SCS1 measures nine concepts, unidimensionality would suggest that they are all related through the underlying trait of CS1 knowledge. This assumption allows us to use more standard IRT models that assume that achievement level is a singular construct that translates monotonically to more correct responses (i.e. a learner higher in the trait will generally answer more questions correctly). If the unidimensionality assumption were not valid, then more complex multidimensional IRT (mIRT) would be required [9].

To test whether a unidimensional IRT model was an acceptable for the SCS1, we conducted a *confirmatory factor analysis* (CFA). CFA, tests whether the relationship observed between variables in the collected data fits a predefined relationship. Because the goal is to see whether a unidimensional model is acceptable, each of the 27 SCS1 questions was specified to load onto a *single* latent variable, CS1 knowledge. The hypothesis was that a single underlying factor will explain sufficient observed variation in all questions.

The CFA was fit in R using the `cfa` function [28]; because test responses are dichotomous, we chose diagonally weighted least squares (WLSMV) as the estimation method. This method assumed that the underlying factor is normally distributed, but does not assume that the observed variables are [19]. To make the model identified, the variance of the latent factor was constrained to 1, so that factor loadings could be estimated for all questions.

Model fit appeared good. The model χ^2 tested the hypothesis that the model fits the data perfectly. $\chi^2(324) = 356.01$ (p=0.11), so we failed to reject that hypothesis. The comparative fit index (CFI), which compares the fit of our model to a null model, was 0.98 (> .90 is acceptable). The root mean square error of approximation (RMSEA) measures model fit with a penalty for more complex models. We found RMSEA = .014, which is acceptable ($\leq$ 0.1). Finally, the standardized root mean square residual (SRMR) is based on the overall difference between the observed and estimated correlation matrices. We found SRMR = .079, which is also acceptable ($\leq$ 0.8). These values all indicate acceptable model fit [14, 30], meaning one factor likely explains variation in these test items.

Table 1 shows the standardized factor loadings for each item, which can be interpreted as regression coefficients. For example, squaring the loading of Q3 (0.58) shows that CS1 knowledge explains $\approx$ 34% of the variation in scores for Q3. We found that a majority of the items had weak to moderate direct effects from the factor. Q20, 24, and 27 all showed poor loading on the factor. In addition to being low magnitude, each of those loadings were not found to be significantly different from zero. These particular items did not appear to be related to the factor of CS1 knowledge.

Table 1 also shows the change in Cronbach's α for each item. Cronbach's α is a measure of internal-consistency reliability. We found that $\alpha = 0.700$, an acceptable level given the SCS1 is primarily used for research and is not a high-stakes test [18, 25]. The table shows the change to α if a question is removed, which we would expect to be negative. If α actually *increased*, that would suggest that question is potentially problematic as the internal consistency of the SCS1 improved. We found that α increased when the same 3 questions which had poor factor loadings (20, 24, 27) were removed.

The residual covariance matrix also suggested that questions 20, 24, and 27 fit poorly. So, we removed these questions.

After dropping these 3 questions and re-running CFA, the α improved to 0.723. Omitting these questions also resulted in a better fitting model, $\chi^2(252) = 271.9$, p=0.19, CFI = 0.99, RMSEA = .013, and SRMR = .076. These fit indices suggest that a single factor of CS1 knowledge sufficiently explained variation in the SCS1 questions.

Table 1: Standardized factor loadings and change in α levels for SCS1 questions. High loading is ideal, suggesting a strong association between a question and the underlying factor (CS1 knowledge). α should be non-increasing.

Num	Factor loading (w/ std. error)	Change in α	Num	Factor loading (w/ std. error)	Change in α
1	0.49 (0.06)	-0.02	15	0.31 (0.08)	0.00
2	0.36 (0.06)	-0.01	16	0.68 (0.06)	-0.02
3	0.58 (0.05)	-0.02	17	0.46 (0.07)	-0.01
4	0.43 (0.08)	-0.01	18	0.16 (0.08)	0.00
5	0.29 (0.09)	0.00	19	0.68 (0.05)	-0.02
6	0.35 (0.06)	-0.01	20**	0.07 (0.08)	+0.01
7	0.46 (0.08)	-0.01	21	0.29 (0.07)	0.00
8	0.39 (0.07)	-0.01	22	0.35 (0.07)	-0.01
9	0.55 (0.06)	-0.02	23	0.43 (0.06)	-0.01
10	0.41 (0.06)	-0.01	24**	0.04 (0.08)	+0.01
11	0.38 (0.07)	-0.01	25	0.28 (0.06)	0.00
12	0.54 (0.06)	-0.02	26	0.43 (0.07)	-0.01
13	0.24 (0.08)	0.00	27**	-0.05 (0.07)	+0.01
14	0.47 (0.06)	-0.01			

** denotes a problematic question dropped from our analysis

4.2 IRT Model Fitting

After verifying the IRT assumptions of unidimensionality and local independence, we fit the response data to models. To evaluate question difficulty and discrimination, we fit the data to dichotomous IRT models with increasing number of parameters. More complex models tend to fit data better, but tend to be less generalizable. Therefore, we choose to fit four common models beginning with the simplest. We fit the Rasch, 1 Parameter Logistic (1PL), 2 Parameter Logistic (2PL), and 3 Parameter Logistic (3PL) models [8] to the data with `question.fit` function from the ltm package in R [27]. We assessed model performance using Akaike information criterion (AIC) and Bayesian information criterion (BIC) [15]. Information criteria transform the log-likelihood of a model by penalizing more complex models (models with more estimated parameters). These criteria can be used to compare similar models to ensure the best fit with the simplest model. We used a χ^2 test to check question fit.

Table 2 compares model performance. We selected the 2PL model because all 24 questions fit. We found it unusual that the 2PL model reported a greater BIC than the less complex Rasch and 1PL models, but we decided having all the questions fitting was more important.

Table 2: Dichotomous model performance. 2PL fit best.

Model	AIC	BIC	Questions that do not fit model
rasch	13104.7	13205.3	3, 19
1PL	13063.7	13168.5	3, 9, 16, 19, 23
2PL	**13018.3**	**13219.6**	**(all questions fit)**
3PL	12974.2	13276.1	3, 6, 22

For 2PL, the probability of learner x getting question j correct given a learner's knowledge level θ, question difficulty α_j, and question discrimination δ_j is $p(x_j = 1|\theta, \alpha_j, \delta_j) = \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}}$.

While we present results for pre-CS1 and pre-CS2 students together, we also used CTT to ensure that any issues we identified were consistent between the two groups. We used CTT measures because did not have appropriate sample sizes to model pre-CS1 and pre-CS2 students separately with IRT.

Table 3: Question parameters for 2PL model. Questions should have $|\delta_j| \leq 3$ and $\alpha_j > 0$, with ** denoting issues.

j	δ_j (SE)	α_j (SE)	j	δ_j (SE)	α_j (SE)
1	1.16 (0.2)	0.91 (0.15)	13**	**3.99 (1.34)**	**0.40 (0.14)**
2	0.89 (0.23)	0.63 (0.13)	14	1.13 (0.21)	0.82 (0.14)
3	0.27 (0.1)	1.26 (0.19)	15**	**3.11 (0.81)**	0.51 (0.14)
4	2.26 (0.42)	0.73 (0.15)	16	1.28 (0.16)	1.35 (0.2)
5**	**3.91 (1.13)**	0.49 (0.15)	17	1.68 (0.29)	0.81 (0.15)
6	0.99 (0.25)	0.59 (0.13)	18**	**5.07 (2.55)**	**0.25 (0.13)**
7	2.33 (0.41)	0.79 (0.16)	19	0.74 (0.11)	1.53 (0.22)
8	1.60 (0.33)	0.66 (0.13)	21	1.58 (0.44)	0.47 (0.12)
9	1.16 (0.18)	1.00 (0.16)	22	1.36 (0.33)	0.57 (0.13)
10	1.27 (0.27)	0.68 (0.13)	23	0.08 (0.13)	0.79 (0.14)
11	1.90 (0.39)	0.63 (0.14)	25	0.78 (0.28)	0.47 (0.12)
12	0.63 (0.13)	1.06 (0.17)	26	1.77 (0.31)	0.78 (0.15)

4.3 Difficulty & Discrimination (RQ2)

Table 3 shows the 2PL model parameters and their standard error (SE). The difficulty means that a learner with exactly that knowledge level should have a 50% chance of answering correctly. Because most questions have a positive difficulty (δ_j) and 0 represents the average knowledge, the SCS1 is a difficult test. Question difficulty (and learner knowledge) represents the number of standard deviations from the mean (z-score). So, difficulty should range from -3 to 3, theoretically accounting for > 99.9% of learners' knowledge levels.

We found that Q5, 13, 15, and 18 are potentially too difficult. For these four questions, only 12-21% of pre-CS1 students got them correct, confirming these questions as being difficult for pre-CS1 students. For the pre-CS2 students, the more difficult questions were Q15 (19% got correct) and Q13 (25%), whereas Q5 (30%) and Q18 (31%) were still difficult but perhaps acceptably so.

The discrimination parameter (α_j) tells us how effectively a question differentiates among learners located near that question's difficulty. More specifically, we can interpret the discrimination to say that a greater α_j means that question j provides more information about learners with a knowledge level near the question's difficulty (δ_j), but this information more rapidly decreases for learners further away from the question's difficulty. Visually, a larger discrimination parameter translates to a steeper logit regression line, such as Q19 in Figure 2. Q19 does an excellent job distinguishing between learners around 0.75, but provides almost no information for learners above 2 or below 0 (average).

While a great discrimination parameter is desirable, there is disagreement on which values are "acceptable." For example, de Ayala 2009 defined 0.8-2.5 as an acceptable discrimination parameter range [8], whereas Baker 2001 defined $0.01 - 0.34$ as "very low" and $0.35 - 0.64$ as "low" discrimination [2]. We flagged Q13 ($\alpha_{13} = 0.40$) and Q18 ($\alpha_{18} = 0.25$) as having potentially problematic discrimination because their discrimination parameters were the lowest among all questions. We used point-biserial correlations (a CTT measure of discrimination [1]) to check discrimination for pre-CS1 and pre-CS2 students separately. Both Q13 and Q18 had a low point-biserial correlation (< 0.30) for both pre-CS1 ($r_{pre-CS1,Q13} = 0.10, r_{pre-CS1,Q18} = 0.18$) and pre-CS2 ($r_{pre-CS2,Q13} = 0.20, r_{pre-CS2,Q18} = 0.28$) students. Because these questions have high difficulty and pre-CS2 learners had more knowledge, it followed that the discrimination for pre-CS2 students were slightly higher (but still too low).

4.4 Estimates of learner knowledge (RQ3)

To better contextualize the difficulty of the SCS1, we used our 2PL model to estimate learner knowledge, something that IRT can measure but CTT cannot. Learner knowledge estimates are summary measures of the posterior distribution of CS1 knowledge conditioned on estimates of question difficulty and observed learner performance [8]. We used `ltm::factor.scores()` [27] with the mean of the posterior distribution as our estimator [3].

Figure 3 shows the relationship between learner knowledge and question difficulty. The estimated learner knowledge from our sample is on the left side, with estimated question difficulties on the right, allowing a clear comparison of learner knowledge to question difficulty levels. Learners in our sample had knowledge ranging from -1.73 to 2.93 and the distribution had a positive/right skew, largely because of the 13% of the sample who were pre-CS2.

With the learner knowledge estimates, we have stronger evidence to suggest that Q5, 13, 14, and 18 (noted as problematic in Table 3) are too difficult. This can clearly be seen in Figure 3 because these questions are above even the highest knowledge learner. These questions have an estimated difficulty greater than all 489 learners in the sample. Furthermore, the difficulty parameters of these questions are ≥ 3, and both knowledge and difficulty are z-scores on a normal distribution, so we would expect only 0.13% of the population to have a knowledge level of 3 or greater. So, these questions should undergo review for revision or removal.

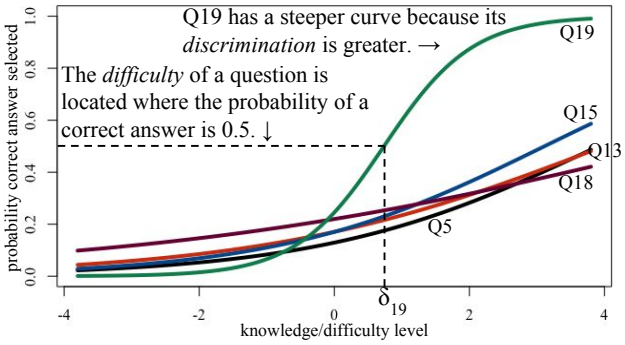

Figure 2: Item characteristic curves for an acceptable question (Q19) and four problematic ones in the SCS1. The flatness of the logistic curves for Q5, 13, 15, and 18 reflect the low/poor discrimination (as reported in Table 3).

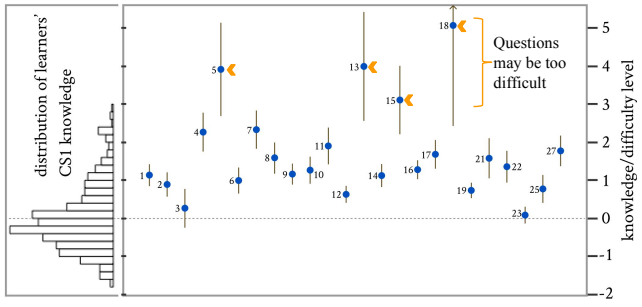

Figure 3: Wright map [10] showing the relationship between learner knowledge (left, ranging from -1.73 to 2.93, mean: 0.00, med.: -0.17) and question difficulty w/ std. error (right).

4.5 Responses for removed questions (RQ4)

Three additional questions that may require revision or removal from the SCS1 are the questions that we dropped prior to running our analysis due to poor factor loadings (Q20, 24, 27; see Table 1). While we dropped these questions because their response patterns deviated from our expectation, factor analysis did not inform us as to *how* these response patterns differed.

Analyzing responses to each incorrect option can provide insight into potential issues relating to test design and learner misconceptions. Each SCS1 question was multiple choice and had 1 correct option and 4 incorrect options, known as *distractors*. Ideally, distractors reflect common misconceptions. But in practice, some options may not represent a common misconception, resulting in the option rarely being selected for learners regardless of knowledge level. We would expect learners of lesser knowledge levels to more frequently select a distractor instead of the correct answer. If learners with greater knowledge levels systematically continue to select a distractor over the correct option, this could suggest confusion relating to the design of the question or a gap in instruction as learners consistently exhibit a misconception. So, we expect the frequency of distractor selection to decrease as knowledge increases.

To better understand how response patterns from dropped questions deviated, we used the IRT-based Nominal Response Model (NRM) [7, 8]. Whereas the 2PL model only considered whether a learner got a question correct or not, the NRM computes the probability of a learner choosing each of the answer options, taking into account the question's difficulty and the learner's knowledge.

In Figure 4, we plotted the probability of learners with varying knowledge levels selecting responses for the three dropped SCS1 questions compared to Q19 (a more ideal question). Each line within a graph represents the probability of a learner selecting a response option as a function of the their knowledge level, with the highest line being the answer choice with the highest probability. For example, for Q19, option C (the correct answer) was the choice with the highest probability for learners with knowledge levels around 0.4 or higher; for learners with knowledge levels around -1.4 or lower, option A was the most probable choice. Q19 displays the expected relationship for a well-functioning question, where learners with more knowledge are more likely to choose the correct answer. In contrast, Q20, 24, and 27 are problematic as they show that learners with more CS1 knowledge (further to the right) are *less* likely to choose the correct answer.

4.6 Expert review of problematic questions

With factor analysis and IRT modeling, we identified specific SCS1 questions that were problematic and may require revision. We said these questions were problematic because they may assess knowledge that is different from the rest of the SCS1 (based on their low factor loadings, Table 1) or too difficult (δ_j too great, Table 3) for the target population of CS1 learners. The next step would be to review these flagged questions to understand exactly what about these questions makes them problematic. One way to do this is with expert review [1, 8]. To conduct expert review, somebody with Educational Psychology and domain-specific expertise (about CS1, in this case) reviews questions that a validation study identified as potentially problematic. For this paper, the authors conducted the

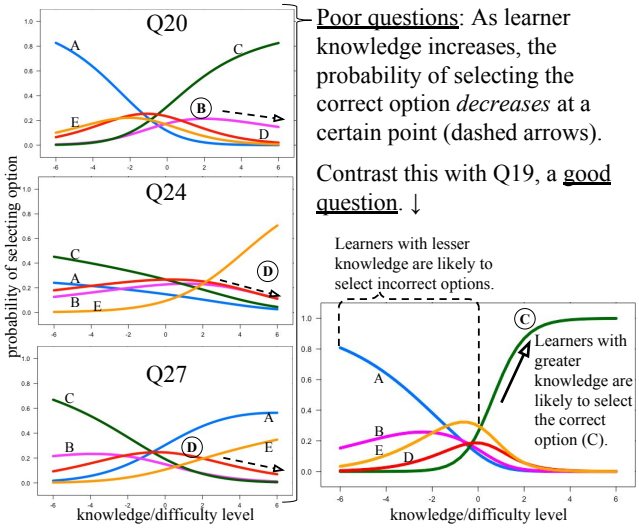

Figure 4: Item characteristic plots for NRM model for dropped questions (Q20, 24, 27) and a good question (Q19) for contrast. Plots show the relationship between learner knowledge and responses. Correct responses are circled.

expert review. In this section, we provide a more detailed explanation of our findings relating to 2 questions which we dropped from the analysis. To maintain the integrity of the SCS1, we were unable to show any of its questions (see Section 6 to access SCS1).

When reviewing the dropped questions (Q20, Q24, Q27), we first identified patterns across all 3 questions. The questions assessed knowledge of function scope (Q20, 27) and recursion (Q24). Prior work has found that both concepts are challenging for CS1 students [17, 32]. Furthermore, scope and recursion have more prominent roles in some programming paradigms than others. For example, almost all pre-CS2 students in our sample took a CS1 course that taught programming in R for working with data; within that curriculum, recursion was not taught and scope was only briefly covered.

We focused our expert review on the questions relating to scope. We looked at the SCS1 to contextualize the patterns in distractor selection for learners of varying knowledge levels (Figure 4). We found that for both questions, learners with below average knowledge (< 0) would select a distractor suggesting a misconception on code tracing; learners with above average knowledge (> 0) would select the distractor which reflected tracing the code correctly but lacking knowledge relating to scope. These similarities in response pattern suggest that these questions may assess similar knowledge and we may be able to drop one of them or revise one to assess different misconceptions relating to scope. Furthermore, the poor factor loading between these questions and the rest of the SCS1 suggest that this knowledge may be somewhat different from other CS1 concepts, or may not be covered sufficiently in CS1 courses.

Further expert review would involve examining the rest of the questions flagged as problematic in Table 3. Because of expert blind spot [22], expert review often includes cognitive interviews with the target population. Then, we can keep, revise, or remove questions.

5 DISCUSSION

In this paper, we conducted a validation study of the SCS1 using IRT to identify differences in question difficulty, discrimination, and response patterns relative to learner knowledge. From this, we found that 3 SCS1 questions may assess knowledge that is different from the rest of the test, and 4 other questions were too difficult for all 489 students in our sample, and potentially too difficult for the target population of CS1 learners. For the 3 questions that may assess different knowledge, we looked at which options were chosen, which further confirmed that the knowledge those questions measured may have been different. Finally, we conducted an expert review to better understand differences in learners' errors.

One way to interpret these results is that they are useful in identifying potentially problematic questions so we can improve the SCS1. In the previous section, we provided an example of how we could do that with expert review of flagged questions. This revealed that, at least in the populations we studied, scope and recursion may not have been covered in their prior coursework. It also revealed opportunity to drop or revise 1 question because it is redundant. However, it is important to conduct cognitive interviews with the target population to verify results from expert review.

Another way to interpret these results is that similar results could have been found through less complex means. Indeed, a limitation to IRT is the complexity of the analysis and requirements for a large sample size. But while CTT might have revealed *that* the test was hard, IRT revealed *how* and *for whom* each question was difficult. In particular, IRT estimated question difficulty and learner knowledge separately so that we could better distinguish whether a question was truly difficult or we happened to have learners with low knowledge levels. With this, we can better justify whether a question is appropriate for a given purpose or certain target population. We found that some items of the SCS1 were too difficult for the population of CS1 learners. Furthermore, there are additional IRT techniques that can help identify whether questions work differently for learners at the same knowledge level but with different characteristics (e.g. males and females) [21], as well as methods which can model tests that measure multiple knowledge constructs [9]. Many aspects of IRT can be useful in future work.

Beyond demonstrating the potential value of IRT, our results have practical implications for research and teaching. First, as with any instrument, the SCS1 needs more refinement to improve the validity of its measurements. Our results suggest that prior work using it as a pre-test may have risked encountering floor effects because of the test's difficulty. At its current level of difficulty, the SCS1 may be more appropriate only as a post-test. This suggests the need for additional instruments that help us measure prior knowledge more reliably, especially as more diverse learners, with more varied prior knowledge, begin to engage in computing education.

6 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1703304, 1735123, 1314399, 1639576, 1829590, and 12566082. Archive of full results can be accessed at github.com/codeandcognition/archive-2019sigcse-xie. We thank Miranda Parker and Mark Guzdial for collecting, sharing, and helping interpret the data. To access the SCS1, contact authors of [26].

REFERENCES

[1] Mary J. Allen and Wendy M. Yen. 2001. *Introduction to Measurement Theory.* Waveland Press.

[2] Frank B. Baker. 2001. *The Basics of Item Response Theory. Second Edition.* ERIC Clearinghouse on Assessment and Evaluation.

[3] R Darrell Bock and Robert J Mislevy. 1982. Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement* 6, 4 (1982), 431–444.

[4] Peter L. Bonate. 2000. *Analysis of Pretest-Posttest Designs.* CRC Press.

[5] Dennis Bouvier, Ellie Lovellette, John Matta, Bedour Alshaigy, Brett A. Becker, Michelle Craig, Jana Jackova, Robert McCartney, Kate Sanders, and Mark Zarb. 2016. Novice Programmers and the Problem Description Effect. In *Proceedings of the 2016 ITiCSE Working Group Reports (ITiCSE '16).* ACM, New York, NY, USA, 103–118. https://doi.org/10.1145/3024906.3024912

[6] Michelle Craig, Jacqueline Smith, and Andrew Petersen. 2017. Familiar Contexts and the Difficulty of Programming Problems. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research (Koli Calling '17).* ACM, New York, NY, USA, 123–127. https://doi.org/10.1145/3141880.3141898

[7] R. Darrell Bock. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 1 (01 Mar 1972), 29–51. https://doi.org/10.1007/BF02291411

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

[9] Johannes Hartig and Jana Höhler. 2009. Multidimensional IRT Models for the Assessment of Competencies. *Studies in Educational Evaluation* 35, 2-3 (June 2009), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

[10] David Torres Irribarra and Rebecca Freund. 2016. IRT Item-Person Map with 'ConQuest' Integration.

[11] Natalie Jorion, Brian D. Gane, Katie James, Lianne Schroeder, Louis V. DiBello, and James W. Pellegrino. 2015. An Analytic Framework for Evaluating the Validity of Concept Inventory Claims. *Journal of Engineering Education* 104, 4 (Oct. 2015), 454–496. https://doi.org/10.1002/jee.20104

[12] David Joyner. 2018. Toward CS1 at Scale: Building and Testing a MOOC-for-Credit Candidate. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale - L@S '18.* ACM Press, London, United Kingdom, 1–10. https://doi.org/10.1145/3231644.3231665

[13] Michael Kane. 2009. Articulating a Validity Argument. In *The Routledge Handbook of Language Testing.* Routledge. https://doi.org/10.4324/9780203181287.ch2

[14] R.B. Kline. 2011. *Principles and Practice of Structural Equation Modeling, Third Edition.* Guilford Publications.

[15] Sadanori Konishi and G. Kitagawa. 2008. *Information Criteria and Statistical Modeling.* Springer, New York.

[16] K Kotovsky, J.R Hayes, and H.A Simon. 1985. Why Are Some Problems Hard? Evidence from Tower of Hanoi. *Cognitive Psychology* 17, 2 (April 1985), 248–294. https://doi.org/10.1016/0010-0285(85)90009-X

[17] Amruth N. Kumar. 2001. Learning the Interaction Between Pointers and Scope in C++. *SIGCSE Bull.* 33, 3 (June 2001), 45–48. https://doi.org/10.1145/507758.377466

[18] Charles E. Lance, Marcus M. Butts, and Lawrence C. Michels. 2006. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods* 9, 2 (April 2006), 202–220. https://doi.org/10.1177/1094428105284919

[19] Cheng-Hsien Li. 2016. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods* 48, 3 (01 Sept. 2016), 936–949. https://doi.org/10.3758/s13428-015-0619-7

[20] Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, Inc.

[21] Patrícia Martinková, Adéla Drabinová, Yuan-Ling Liaw, Elizabeth A. Sanders, Jenny L. McFarland, and Rebecca M. Price. 2017. Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *Cell Biology Education* 16, 2 (2017), rm2. https://doi.org/10.1187/cbe.16-10-0307

[22] Mitchell J Nathan, Kenneth R Koedinger, and Martha W Alibali. 2001. Expert Blind Spot: When Content Knowledge Eclipses Pedagogical Content Knowledge. *Proceeding of the Third International Conference on Cognitive Science* (2001), 6.

[23] Greg L. Nelson, Benjamin Xie, and Amy J. Ko. 2017. Comprehension First: Evaluating a Novel Pedagogy and Tutoring System for Program Tracing in CS1. In *2017 ACM Int'l Computing Education Research Conf. (ICER '17).* ACM, 2–11.

[24] Melvin R Novick. 1966. The axioms and principal results of classical test theory. *Journal of mathematical psychology* 3, 1 (1966), 1–18.

[25] Jum C. Nunnally. 1978. *Psychometric Theory* (2d ed. ed.). McGraw-Hill, New York.

[26] Miranda C. Parker, Mark Guzdial, and Shelly Engleman. 2016. Replication, Validation, and Use of a Language Independent CS1 Knowledge Assessment. In *2016 ACM Int'l Computing Education Research Conf. (ICER '16).* ACM, 93–101.

[27] Dimitris Rizopoulos. 2018. Latent Trait Models under IRT.

[28] Yves Rosseel. 2018. *The Lavaan Tutorial.* Technical Report. Ghent University.

[29] Marko Sarstedt and Erik Mooi. 2014. *Cluster Analysis.* Springer Berlin Heidelberg, Berlin, Heidelberg, 273–324. https://doi.org/10.1007/978-3-642-53965-7_9

[30] James B. Schreiber, Amaury Nora, Frances K. Stage, Elizabeth A. Barlow, and Jamie King. 2006. Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research* 99, 6 (2006), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

[31] A.H. Segars. 1997. Assessing the Unidimensionality of Measurement: A Paradigm and Illustration within the Context of Information Systems Research. *Omega* 25, 1 (Feb. 1997), 107–121. https://doi.org/10.1016/S0305-0483(96)00051-5

[32] Raja Sooriamurthi. 2001. Problems in Comprehending Recursion and Suggested Solutions. In *Proceedings of the 6th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '01).* ACM, New York, NY, USA, 25–28. https://doi.org/10.1145/377435.377458

[33] Allison Elliott Tew. 2010. *Assessing Fundamental Introductory Computing Concept Knowledge in a Language Independent Manner.* Ph.D. Dissertation. Georgia Institute of Technology.

[34] A. E. Tew and B. Dorn. 2013. The Case for Validated Tools in Computer Science Education Research. *Computer* 46, 9 (Sept. 2013), 60–66. https://doi.org/10.1109/MC.2013.259

[35] Allison Elliott Tew and Mark Guzdial. 2011. The FCS1: A Language Independent Assessment of CS1 Knowledge. In *Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education (SIGCSE '11).* ACM, New York, NY, USA, 111–116. https://doi.org/10.1145/1953163.1953200

[36] Dion Timmermann and Christian Kautz. 2016. Design of Open Educational Resources for a Programming Course with a Focus on Conceptual Understanding. In *Proceedings of the 44th SEFI Annual Conference.* Tampere, Finland.

[37] D. Timmermann, C. Kautz, and V. Skwarek. 2016. Evidence-Based Re-Design of an Introductory Course "Programming in C". In *2016 IEEE Frontiers in Education Conference (FIE).* 1–5. https://doi.org/10.1109/FIE.2016.7757492

[38] Colin S. Wallace and Janelle M. Bailey. 2010. Do Concept Inventories Actually Measure Anything? *Astronomy Education Review* 9, 1 (Dec. 2010). https://doi.org/10.3847/AER2010024