Supplementary Info for *An Item Response Theory Evaluation of a Language-Independent CS1 Knowledge Assessment* (Xie et al. SIGCSE 2019)

Prepared by Benjamin Xie (bxie@uw.edu), Sep 2019.

Paper citation:
Benjamin Xie, Matthew J. Davidson, Min Li, and Andrew 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). ACM, New York, NY, USA, 699-705. DOI: https://doi.org/10.1145/3287324.3287370

Files

- `data_sample`: directory of sample data **to reflect structure of input data**. Running IRT analysis on this will fail because the sample size is too small and the "answers" are made up anyways.
    - `scs1_with_demographics.csv`: data from institution 1 with columns of demographic data
    - `SCS1_PreTest_OnlineCS1.csv`: data from institution 2 for pre-test
    - `SCS1_PostTest_OnlineCS1.csv`: data from institution 2 for pre-test (same structure as SCS1_PreTest_OnlineCS1.csv)
    - `scs1_answers_FAKE.csv`: fake SCS1 "answers" (made up via random number generator). This is to protect the integrity of the instrument
- `response_plots`: plots for distractor analysis (distribution of responses to each option).
    - `nrm_item01` - `nrm_item27`: Nominal Response Model (NRM) plots of each item. **Note that the NRM plots may differ from the ones in the paper, blog post, and/or presentation** (see explanation below).
    - `ctt_responses`: histogram of responses for each item
- `scs1_analysis.Rmd`: R markdown file where all analysis was done (data cleaning, factor analysis, CTT, IRT, generating figures). Note that this file was modified to
    - 1) replace links to real data with sample data files which are equivalent in column structure
    - 2) reflect a bug fix in the package we used for generating the item characteristic plots fo the Nominal Response Model (Fig 4 in paper). See below for more info.

Potential Issues with NRM plots

In brief, the authors believe the NRM plots in the paper (Figure 4) and included in this supplementary information are *correct*, but the one used in the presentation to reflect a poor item was incorrect. Here's a brief explanation as we understand it:

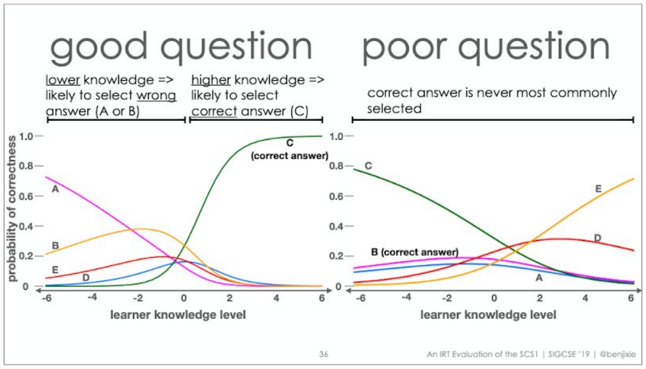
Our NRM plots were made with the `mirt()` function in the mirt R package (https://cran.r-project.org/web/packages/mirt/mirt.pdf). When we first wrote code to conduct this analysis around Feb 2018, we experienced a bug where plots were flipped along the y-axis ("reflected"). We verified this bug by comparing with CTT distractor distribution and 2PL item parameters. Sometime around when we were preparing our paper presentation (Jan 2019), we installed an updated version of mirt which likely fixed this bug. But we did not know this, so we manually reflected the figure again in the presentation. we do As of Sep 2019, I believe this bug has been fixed.

See figures on next page.

So main takeaways:

- Despite our mistake in the presentation, we believe the paper still accurately reflects our analysis.
- Sanity check your IRT results with simple and interpretable CTT results.

Incorrect slide from presentation: plot for "poor question" (item 20) is backwards (reflected across y-axis).



good question

lower knowledge => likely to select wrong answer (A or B)

higher knowledge => likely to select correct answer (C)

poor question

correct answer is never most commonly selected

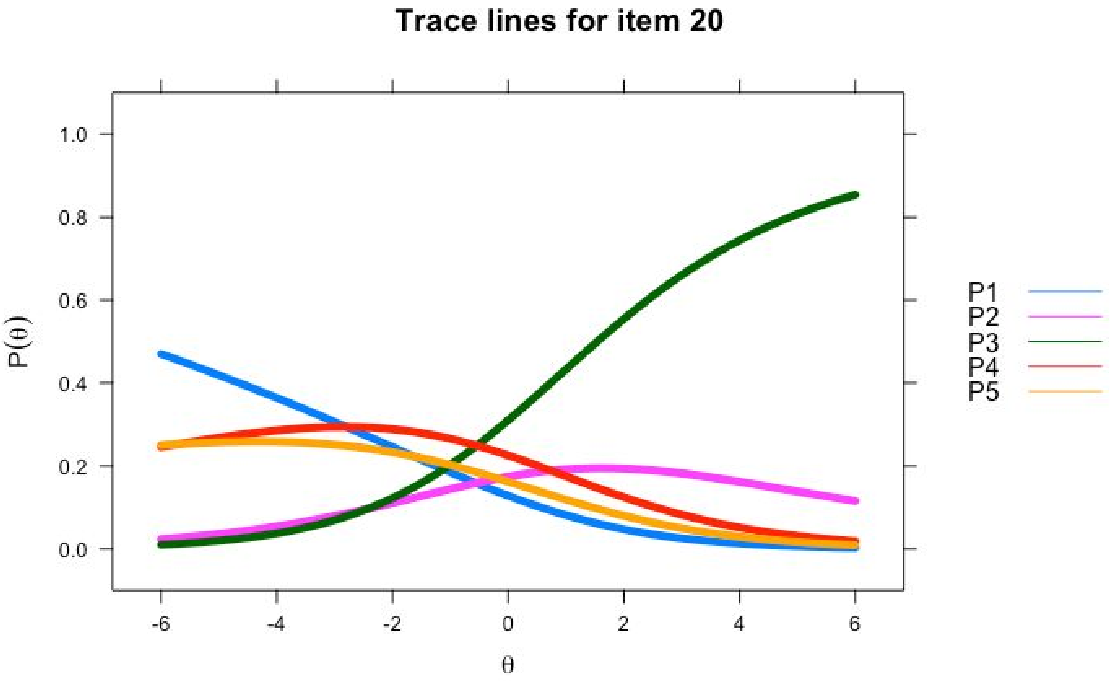
36

An IRT Evaluation of the SCS1 | SIGCSE '19 | @benjixie

this shows us the relationship between a learner's CS1 knowledge and which answer choice they selected. lets us diagnose that something is going wrong with Q20, because the correct answer, B, is less likely to be chosen as ability increases.

By placing learners and question difficulty on the same scale, IRT can help us disentangle specific performance on test questions from properties of the questions themselves.

Correct plot of "poor question" (shown correctly in paper, included in supplementary information).



Factor Loadings

This table on factor loadings of each item was cut for length, so we're including it here for comparison in replication studies:

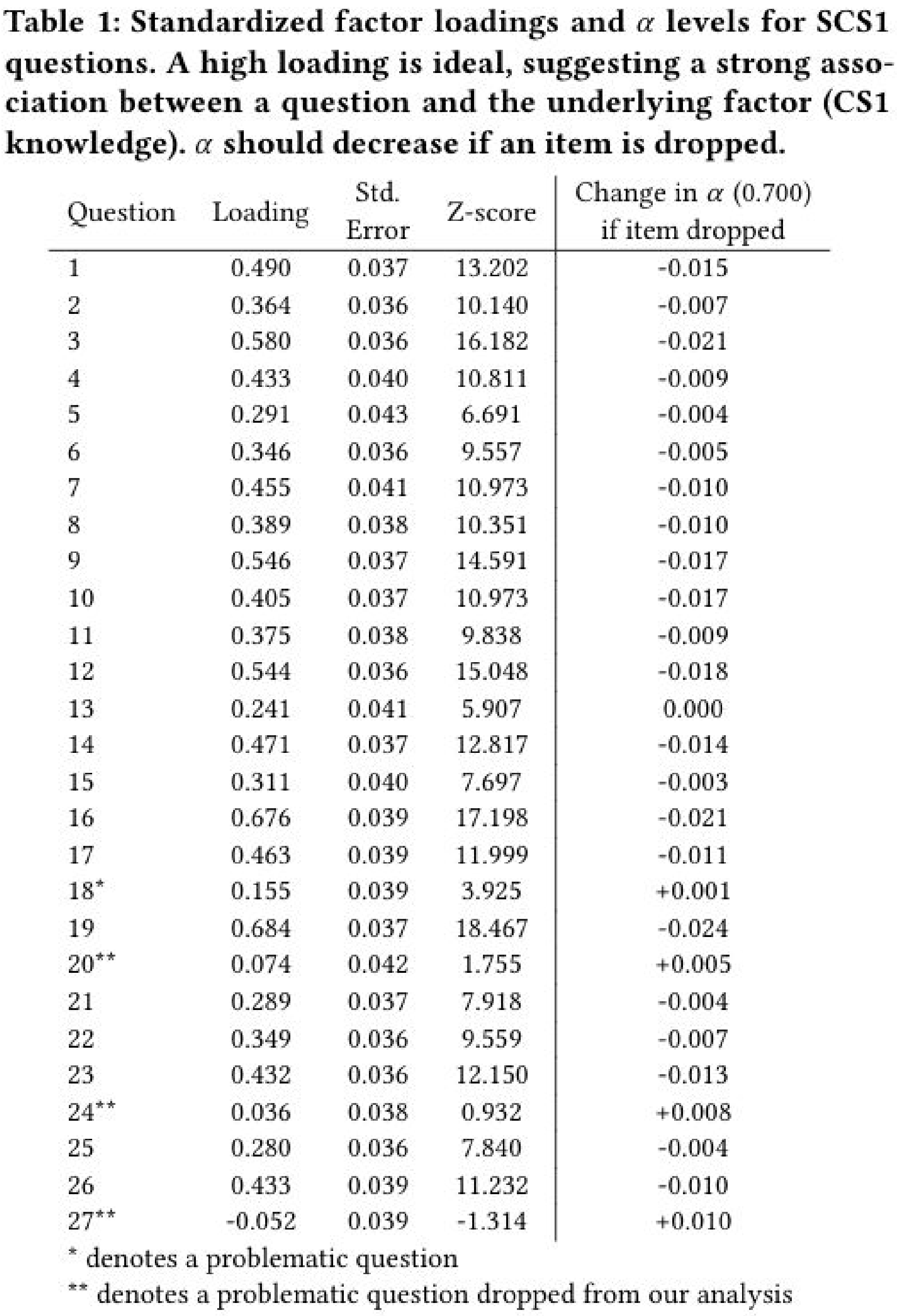


Table 1: Standardized factor loadings and α levels for SCS1 questions. A high loading is ideal, suggesting a strong association between a question and the underlying factor (CS1 knowledge). α should decrease if an item is dropped.

Question	Loading	Std. Error	Z-score	Change in α (0.700) if item dropped
1	0.490	0.037	13.202	-0.015
2	0.364	0.036	10.140	-0.007
3	0.580	0.036	16.182	-0.021
4	0.433	0.040	10.811	-0.009
5	0.291	0.043	6.691	-0.004
6	0.346	0.036	9.557	-0.005
7	0.455	0.041	10.973	-0.010
8	0.389	0.038	10.351	-0.010
9	0.546	0.037	14.591	-0.017
10	0.405	0.037	10.973	-0.017
11	0.375	0.038	9.838	-0.009
12	0.544	0.036	15.048	-0.018
13	0.241	0.041	5.907	0.000
14	0.471	0.037	12.817	-0.014
15	0.311	0.040	7.697	-0.003
16	0.676	0.039	17.198	-0.021
17	0.463	0.039	11.999	-0.011
18*	0.155	0.039	3.925	+0.001
19	0.684	0.037	18.467	-0.024
20**	0.074	0.042	1.755	+0.005
21	0.289	0.037	7.918	-0.004
22	0.349	0.036	9.559	-0.007
23	0.432	0.036	12.150	-0.013
24**	0.036	0.038	0.932	+0.008
25	0.280	0.036	7.840	-0.004
26	0.433	0.039	11.232	-0.010
27**	-0.052	0.039	-1.314	+0.010

* denotes a problematic question

** denotes a problematic question dropped from our analysis

SCS1 is not "language-independent"

The title of this paper describes the instrument we analyzed as a "[programming] language-independent" assessment. We used this terminology to be consistent with how the SCS1 was referred to when it was created. The SCS1 authors now feel "language-independent" is an incorrect label.

Read more:
https://cacm.acm.org/blogs/blog-cacm/238782-we-should-stop-saying-language-independent-we-dont-know-how-to-do-that/fulltext