Assessing the assessment: Mutual information between response choices and factor scores

Partially supported by Grant No. 1611482

Cole Walsh and N.G. Holmes

Laboratory of Atomic & Solid State Physics, Cornell University, Ithaca, NY, 14853

Mutual Information

$$I(F;R) = \sum_{f \in \mathcal{F}} \sum_{r \in \{0,1\}} p(r,f) \log_2 \frac{p(r,f)}{p(r)p(f)}$$

p(r), marginal probability of selecting/not selecting a response choice

p(f), marginal probability of observing a score f on factor F

p(r, f), joint probability of observing r and f for a response choice - factor combination

Quantitative Interpretation

Suppose I know the distribution of factor scores (say, for the *Evaluating Models* factor) and I wanted to guess a particular student's score on that factor. I take the optimal guessing strategy (dividing the probability distribution in half with each guess: "is f greater than f_0?").

I(F; R) is the reduction in the number of yes/no guesses required to exactly guess f after observing R.

Why Mutual Information?

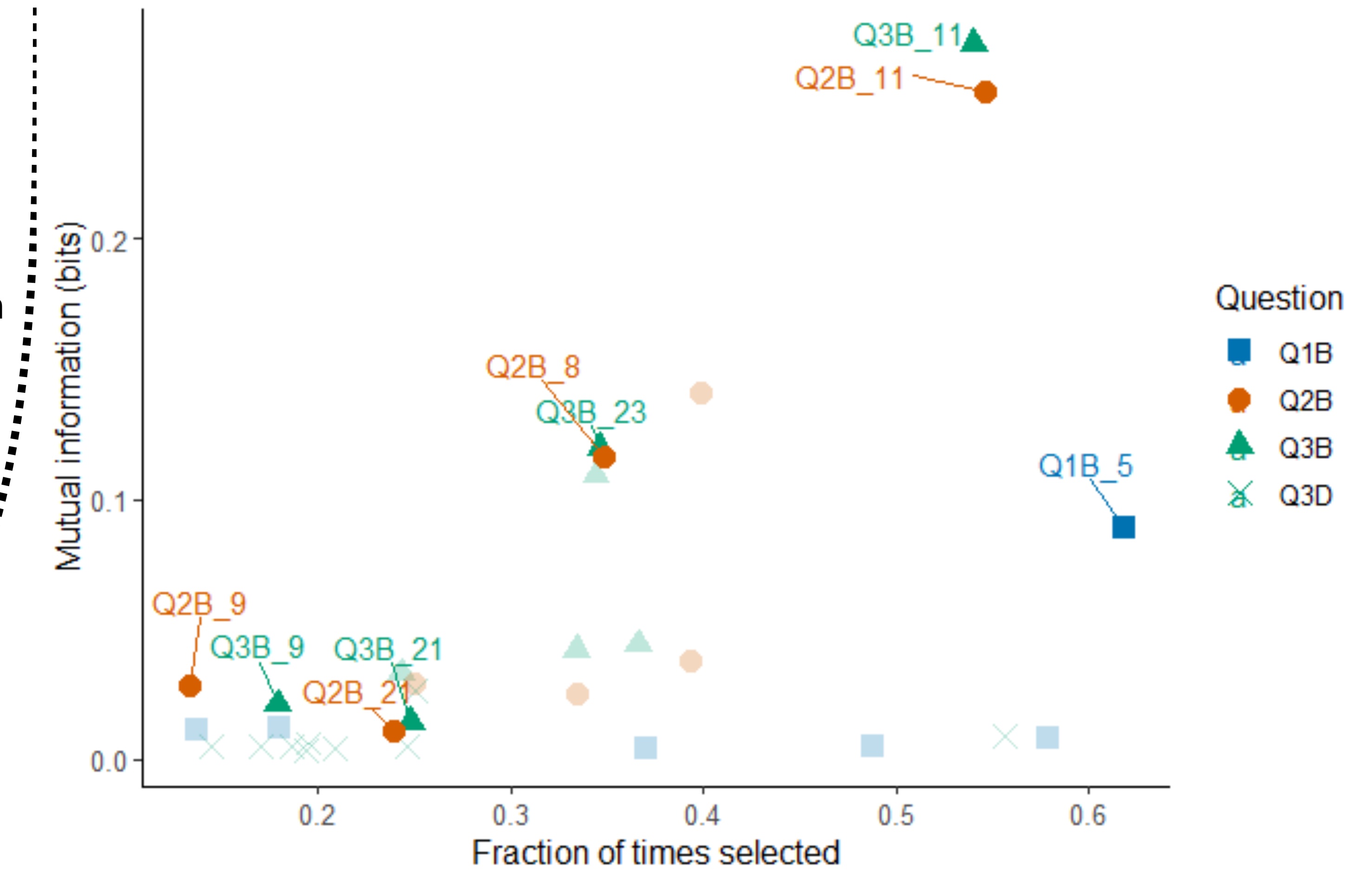
Mutual information can be used alongside more traditional methods, such as classical test theory (CRT) and item response theory (IRT), to evaluate the utility of individual questions. Unlike CRT, and without the model assumptions of IRT, using mutual information we can evaluate the utility of the response choices avaliable to each question based on how much information they provide about latent student abilities.

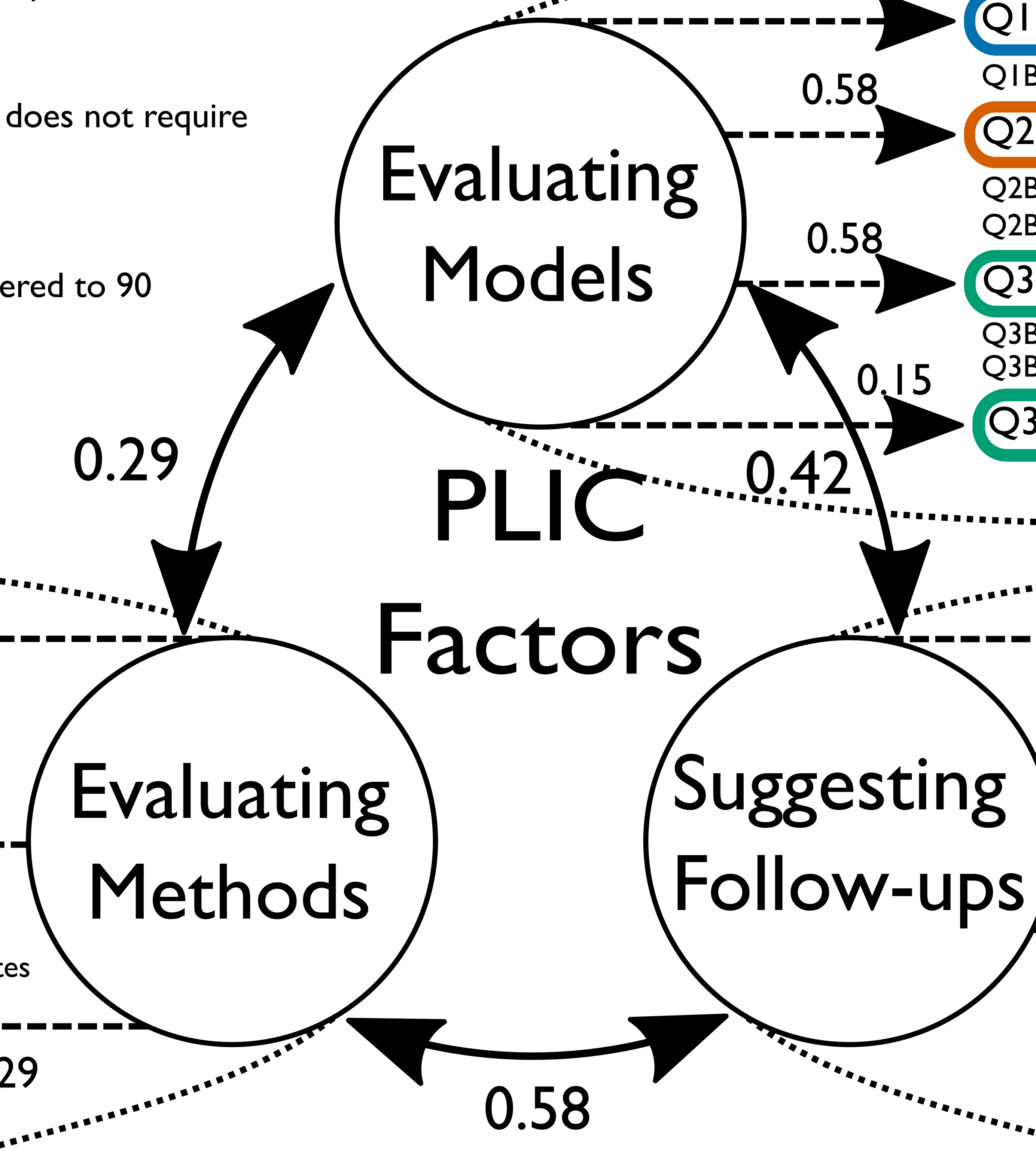
Unlike IRT, calculating mutual information between response choices and factor scores does not require that questions are scored on a binary scale.

Data & Factor Analysis

Data was collected from the Physics Lab Inventory of Critical thinking (PLIC) administered to 90 classes across 39 unique institutions for a total 7525 surveys.

A confirmatory factor analysis (CFA) is performed to evaluate the proposed factor structure. Maximum likelihood estimation is used to extract variances from the data. Factor scores are calculated using Thurstone's regression method.





Q1B: What features were most important in comparing the two k-values?

Q1B_5: the difference in the k-values compared to the uncertainty

Q2B: What features were most important in comparing the fit to the data?

Q2B_8: how close the points are to the line; Q2B_21: the number of points above the line compared to below; Q2B_9: the number of outliers; Q2B_11: how close the points are to the line compared to the uncertainties

Q3B: What features were most important in comparing the NEW fit to the data?

Q3B_9: the number of outliers; Q3B_11: how close the points are to the line compared to the uncertainties; Q3B_23: how close the points are to the line; Q3B_21: the number of points above the line compared to below

Q3D: Which items reflect your reasoning for determining which fit Group 2 should use?

Q1D: What features of Group 1's method were most important for evaluating the method?

Q1D_3: the number of masses tested; Q1D_51: how they tested other variables; Q1D_61: clarity, organization, or detail of their lab notes

Q2D: What features of Group 2's method were most important for evaluating the method?

Q2D_4: the number of masses tested; Q2D_21: how they tested other variables; Q2D_33: their analysis and calculations; Q2D_35: clarity, organization, or detail of their lab notes

Q4B: What features were most important for comparing the two groups?

Q4B_4: the number of masses tested

Q1E: What do you think Group 1 should do next?

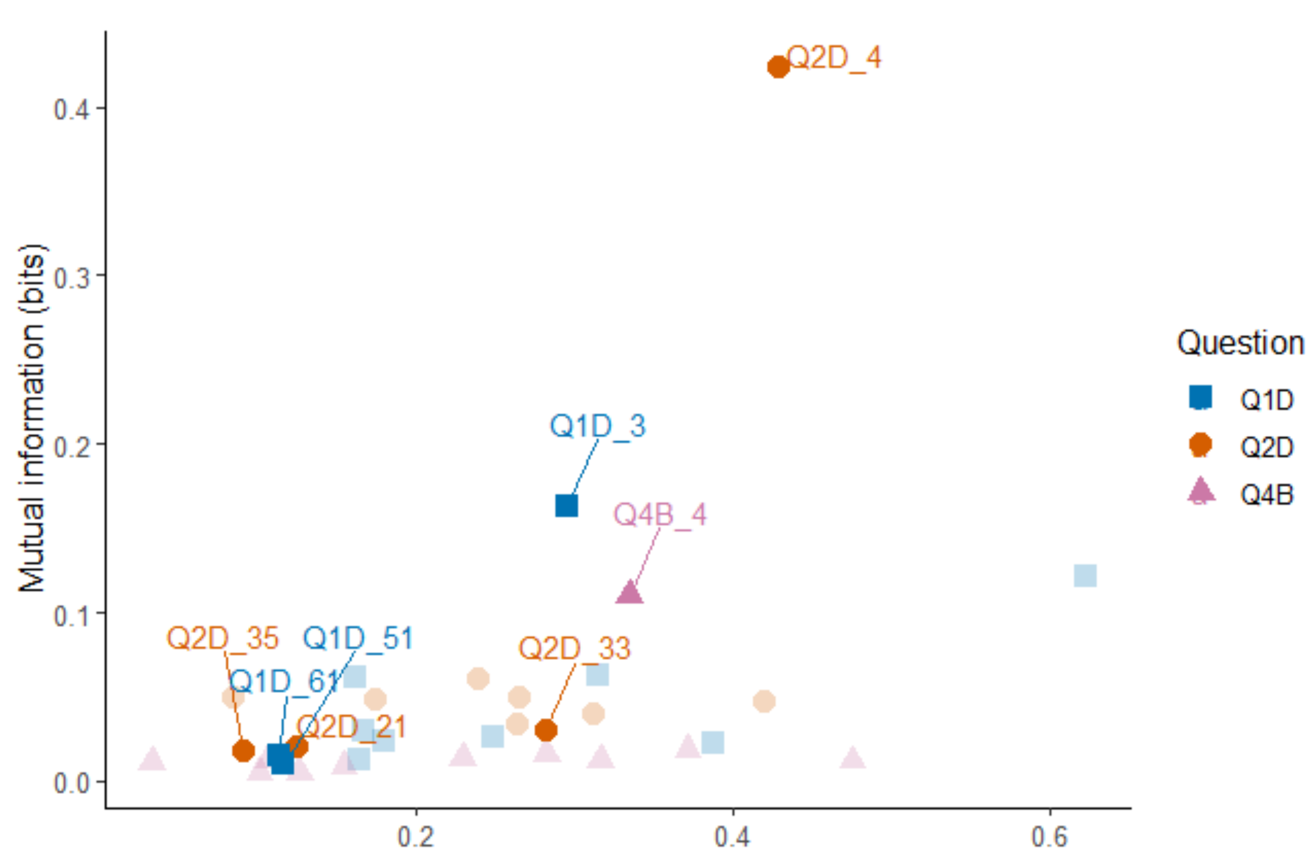
Q1E_1: test more masses; Q1E_4: test other variables; Q1E_13: increase the number of bounces per trial

Q2E: What do you think Group 2 should do next?

Q2E_4: test more masses; Q2E_6: test other variables; Q2E_14: increase the number of bounces per trial

Q3E: What do you think Group 2 should do next? (after changing intercept)

Q3E_11: design a new experiment to test the non-zero intercept; Q3E_13: test other variables; Q3E_14: test more masses; Q3E_20: increase the number of bounces per trial



Takeaways

Results from the CFA indicate the proposed factor structure adequately models the data (CFI > 0.90; RMSEA < 0.05; SRMR < 0.05). Researchers and instructors who use the PLIC can separate students' scores on the instrument into three factor scores and evaluate their data in this context.

Similar to IRT, this method using mutual information between response choices and factor scores allows us to examine which response choices provide the most information about a student's latent abilities.

The most novice and expert response choices (as identified by expert physicists) are typically the most informative. This is not always the case; certain response choices, such as Q2D_33, are expert-like and worth more points, but are relatively uninformative about students' latent abilities — they are picked by high and low performing students.

We can use this method as part of the assessment development process to drop (or modify) relatively uninformative response choices, add new ones, and repeat!