Dimensionality of the Force Concept Inventory: Comparing Bayesian Item Response Models

Xiaowen Liu
Eric Loken
University of Connecticut

Overview

- Force Concept Inventory
- Bayesian implementation of one- and two-dimensional IRT
- Placing IRT results in other contexts
- Summary and future directions

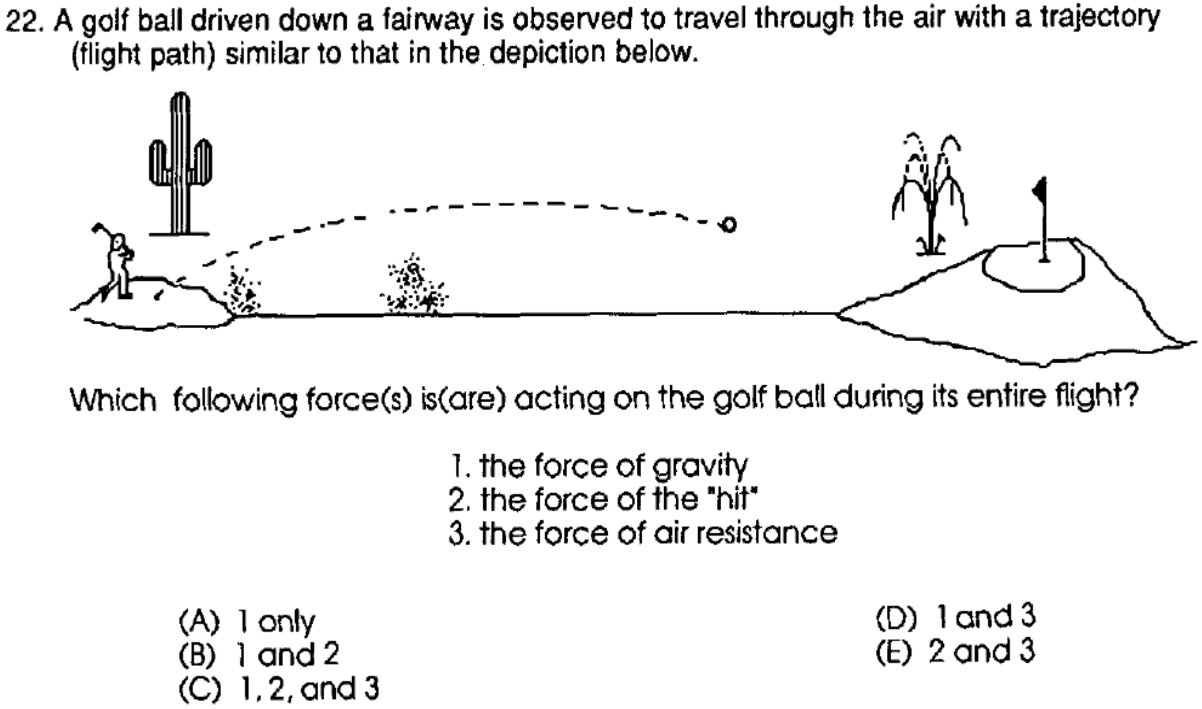
Force Concept Inventory(FCI)

- The Force Concept Inventory (FCI) is very widely used in college physics classes.
- Force Concept Inventory (FCI) was designed as an instrument for probing Newtonian force concepts and commonsense misconceptions. The questions do not require any calculations – just conceptual understanding.
- The core concepts all relate to Newtonian laws of motion, and the questions look very simple on the surface, but require the student to answer based on a true appreciation of Newton’s laws and what they mean, as opposed to just using basic (and incorrect) intuitions about force and motion.

Force Concept Inventory(FCI)

- The test was created with specific misconceptions in mind, and the answer choices reflect this. In fact the distractors are basically as important as the correct answers.
- In this inventory, Newtonian force concepts were decomposed as six conceptual dimensions and each dimension has its own sub-dimension (Hestenes, Wells, and Swackhamer, 1992).

Force Concept Inventory(FCI)—An Example



22. A golf ball driven down a fairway is observed to travel through the air with a trajectory (flight path) similar to that in the depiction below.

Which following force(s) is(are) acting on the golf ball during its entire flight?

1. the force of gravity
2. the force of the "hit"
3. the force of air resistance

(A) 1 only
(B) 1 and 2
(C) 1, 2, and 3
(D) 1 and 3
(E) 2 and 3

Correct Answer: D
Misconception: B, C, E
Impetus supplied by “hit”

Previous Research on Checking Dimensionality of FCI

1. Exploratory factor analysis (EFA)

- Huffman and Heller (1995) found few main factors and lots of ambiguous factors which cannot correspond to the six proposed dimensions.
- Scott, Schumayer, and Gray (2012) identified five factors and interpreted these factors as different Newtonian sub-concepts.
- Semak et. al (2017) identified evolution of response patterns of FCI from pre- and post-tests. They extracted five and six factors in FCI pre-test and post-test respectively.

Previous Research on Checking Dimensionality of FCI

1. Exploratory factor analysis (EFA)

- Issues

  1. Over extraction-- Eigenvalues have similar values except the first one.

  2. Interpretation of extracted factors did not match to the six proposed dimensions. This may imply that test proposed dimensions do not match the factors actually measured.

Previous Research on Checking Dimensionality of FCI

2. Item Response Theory (IRT)

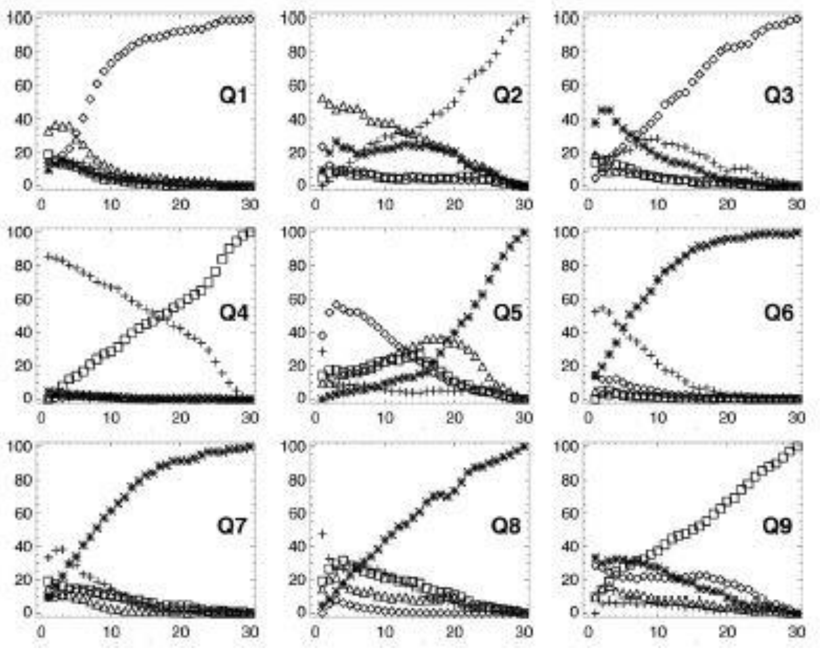
- Rasch model: Planinic, Ivanjek, and Susac (2010) employed Rasch model for item difficulty estimation.
- 3PL:Wang and Bao (2010) applied three-parameter item response model to analyze a college course FCI data.
- Both groups said their analysis supported unidimensionality for FCI.

Previous Research on FCI

3. Item Response Curve (IRC)

Morris et al. (2012) applied Item response curve to FCI to investigate distractor functio of each item at each ability level.

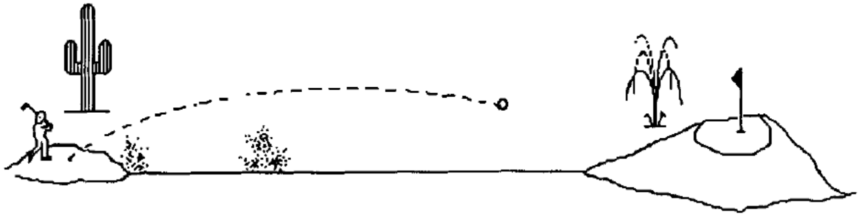
- Nonparametric method
- Describes the proportion of selection for each option in each item.
- They explicated the information provided by misconception alternatives.



Previous Research on FCI

3. Item Response Curve (IRC)

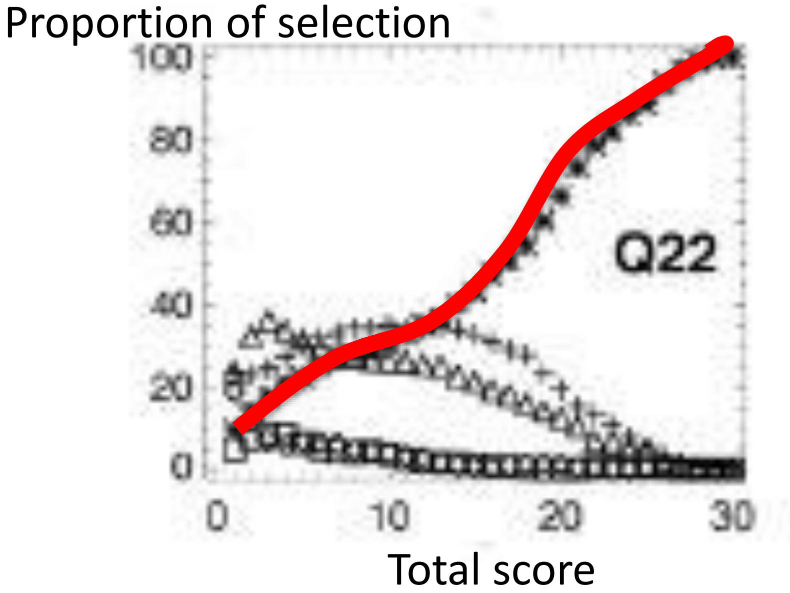
22. A golf ball driven down a fairway is observed to travel through the air with a trajectory (flight path) similar to that in the depiction below.



Which following force(s) is(are) acting on the golf ball during its entire flight?

1. the force of gravity
2. the force of the "hit"
3. the force of air resistance

(A) 1 only
(B) 1 and 2
(C) 1, 2, and 3
(D) 1 and 3
(E) 2 and 3



Bayesian Item Response Theory

- A Bayesian approach may improve the reliability of the statistical inferences by treating parameters as random variables and incorporating prior information.
- The flexibility of the model: define prior distributions for the item response model parameters which can handle more complex sampling designs comprising complex dependency structures (Fox, 2010).

Purpose

In this study, we explored test structure of FCI through Bayesian approach by comparing two models.

Unidimensional two-parameter model

Multidimensional two-parameter model

Bayesian IRT--Model

- Bayesian Two-PL Two-dimension IRT model

$$P(\theta, \xi | y) \propto p(y| \theta, \xi)p(\theta, \xi)$$

θ is the person parameter and ξ is item parameters (item difficulty and discrimination).

$$P(y=1| \theta 1, \theta 2, a1, a2, b)=\frac{1}{1+\exp[-(a_1\theta_1+a_2\theta_2+b)]}$$

- Estimation-- Markov chain Monte Carlo (MCMC)

Bayesian IRT--Prior

	theta	a	b
2p1dim	normal(0,1)	lognormal(0,1)	beta ~ normal(mu_beta,sigma_beta) mu_beta ~ normal(0,5) (hyper) sigma_beta ~ cauchy(0,5) (hyper)
2p2dim	normal(0,1)	alpha1 ~ lognormal(0, 1)	beta ~ normal(mu_beta,sigma_beta) mu_beta ~ normal(0,5) (hyper) sigma_beta ~ cauchy(0,5) (hyper)

Bayesian IRT--Model

- R-STAN was used for fitting the Bayesian IRT models (Luo & Jiao, 2017).
- The large sample (N = 1169) of FCI student responses was collected from a large introductory physics class at the end of the semester.

Bayesian IRT--Result

Item	a	b	a1	a2	b
1	1.10	2.47	1.11	0.00	2.70
2	0.77	0.73	0.66	0.40	0.56
3	0.98	1.29	1.09	0.21	1.31
4	1.46	2.09	0.74	1.50	3.21
5	1.57	0.76	1.05	1.74	1.32
6	1.16	2.52	1.15	0.37	2.93
7	0.90	2.19	0.76	0.47	1.96
8	1.09	0.93	1.19	0.27	1.05
9	0.91	0.61	1.02	0.17	0.58
10	1.85	1.43	1.76	0.77	2.67
11	1.17	1.10	0.70	1.31	1.39
12	1.24	2.06	1.26	0.41	2.59
13	2.66	0.87	1.99	2.08	2.37
14	1.09	0.64	1.26	0.19	0.75
15	0.32	1.85	0.12	0.54	0.60
16	0.83	2.22	0.57	0.61	1.84
17	1.44	0.12	1.14	0.97	0.16
18	1.62	1.00	1.02	1.96	1.87
19	1.34	1.57	1.35	0.48	2.15
20	1.13	1.49	0.99	0.58	1.68
21	0.90	0.09	0.98	0.19	-0.07
22	1.38	0.80	1.55	0.35	1.19
23	1.18	0.36	1.51	0.12	0.49
24	1.44	1.79	1.50	0.47	2.65
25	1.98	0.25	1.69	1.04	0.50
26	2.10	0.04	1.92	0.91	-0.08
27	1.11	1.10	1.38	0.15	1.33
28	1.94	1.60	1.22	1.66	3.19
29	0.65	4.00	0.21	0.57	2.59
30	1.50	0.47	1.09	1.28	0.72

Table 1. Item parameter estimations for the two item response models

Bayesian IRT—Model Fit

Widely available information criterion (WAIC;Watanabe, 2010) and leave-one-out cross-validation (LOO; Vehtari, Gelman, &Gabry, 2016a) are proved to be superior to more traditional methods such as the likelihood ratio test, AIC, BIC, and DIC (Harbi, 2016).

In stan, the posterior distributions can be obtained and used to compute the WAIC and LOO.

	Unidimensional 2PL	Multidimensional 2PL
LOO	31447.7	31186.5
WAIC	31422.3	31109.6

Table 2. Fit indices for comparing the two item response models

Two Ways to Explain the Second Dimension

Item Response Curve

the items with high loadings on the second dimension have strong distractors

K-means clustering

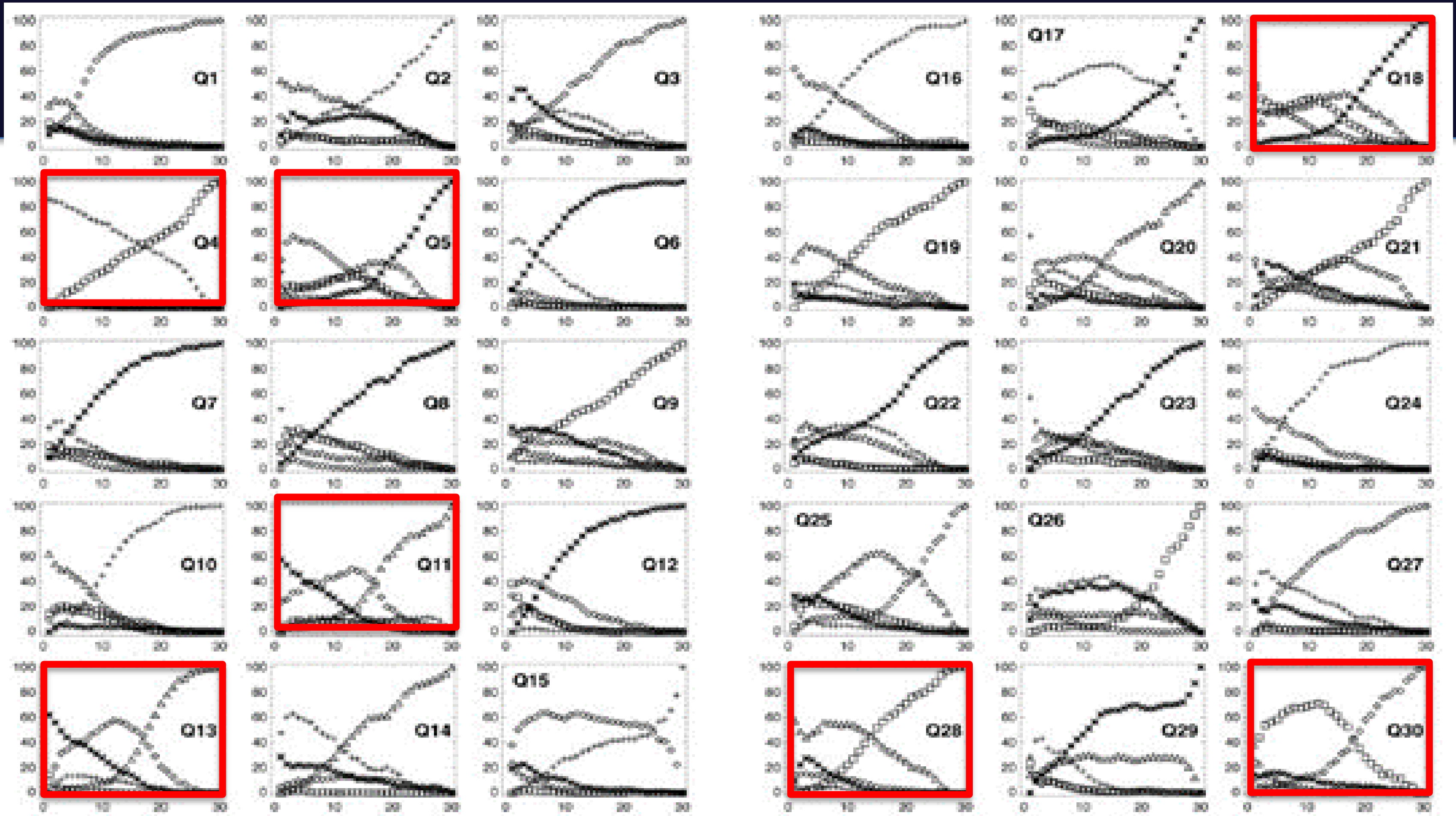
the items with high loadings on the second dimension can differentiate the medium and the weak group.

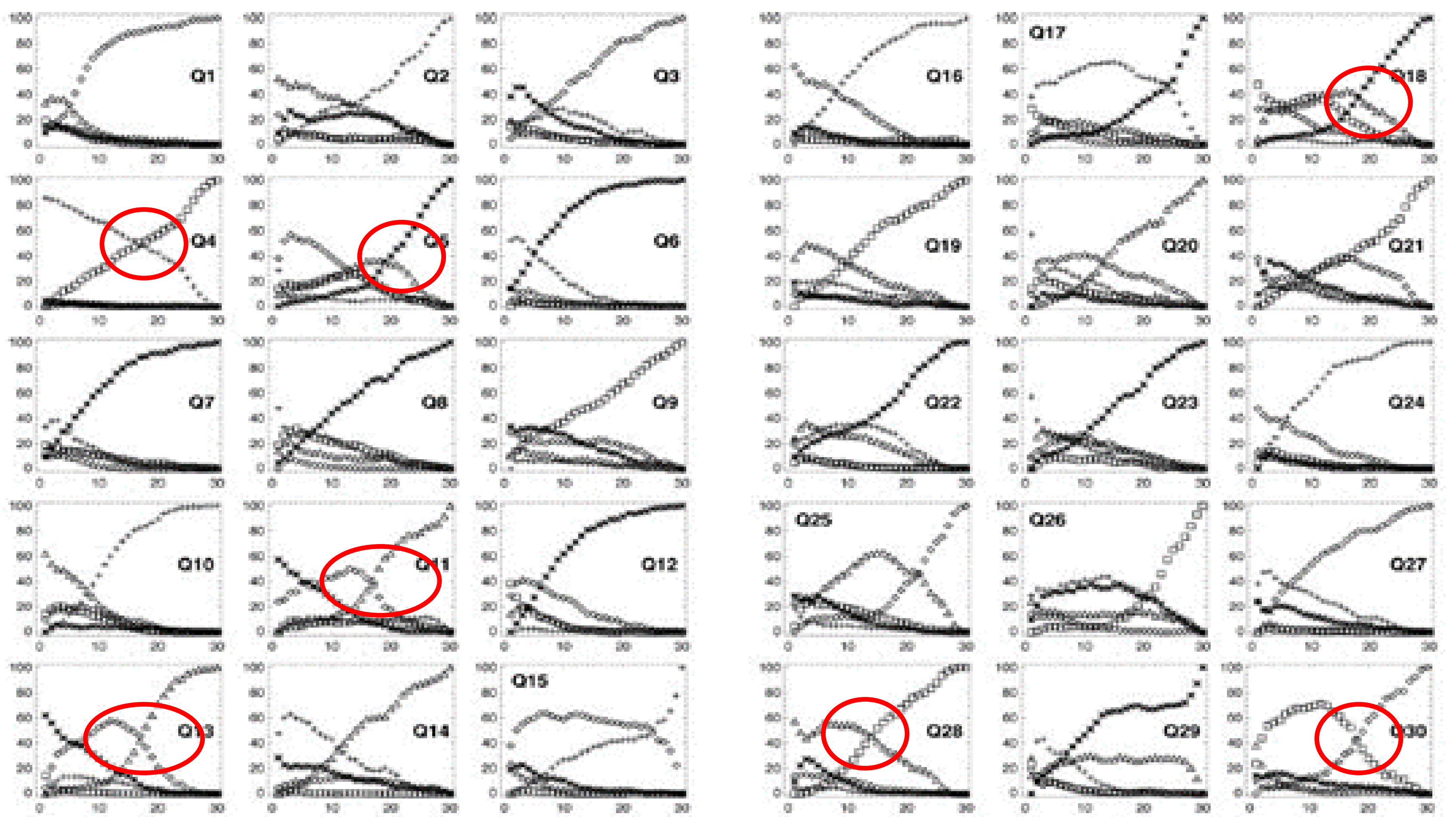
Explain the Second Dimension --Item Response Curve

Item 4, 5, 11, 13, 18, 28, 30

Table 1. Item parameter estimations for the two item response models

Item	a	b	a1	a2	b
1	1.10	2.47	1.11	0.00	2.70
2	0.77	0.73	0.66	0.40	0.56
3	0.98	1.29	1.09	0.21	1.31
4	1.46	2.09	0.74	1.50	3.21
5	1.57	0.76	1.05	1.74	1.32
6	1.16	2.52	1.15	0.37	2.93
7	0.90	2.19	0.76	0.47	1.96
8	1.09	0.93	1.19	0.27	1.05
9	0.91	0.61	1.02	0.17	0.58
10	1.85	1.43	1.76	0.77	2.67
11	1.17	1.10	0.70	1.31	1.39
12	1.24	2.06	1.26	0.41	2.59
13	2.66	0.87	1.99	2.08	2.37
14	1.09	0.64	1.26	0.19	0.75
15	0.32	1.85	0.12	0.54	0.60
16	0.83	2.22	0.57	0.61	1.84
17	1.44	0.12	1.14	0.97	0.16
18	1.62	1.00	1.02	1.96	1.87
19	1.34	1.57	1.35	0.48	2.15
20	1.13	1.49	0.99	0.58	1.68
21	0.90	0.09	0.98	0.19	-0.07
22	1.38	0.80	1.55	0.35	1.19
23	1.18	0.36	1.51	0.12	0.49
24	1.44	1.79	1.50	0.47	2.65
25	1.98	0.25	1.69	1.04	0.50
26	2.10	0.04	1.92	0.91	-0.08
27	1.11	1.10	1.38	0.15	1.33
28	1.94	1.60	1.22	1.66	3.19
29	0.65	4.00	0.21	0.57	2.59
30	1.50	0.47	1.09	1.28	0.72


Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q18
Q19
Q20
Q21
Q22
Q23
Q24
Q25
Q26
Q27
Q28
Q29
Q30
0
20
40
60
80
100
0
10
20
30


Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q18
Q19
Q20
Q21
Q22
Q23
Q24
Q25
Q26
Q27
Q28
Q29
Q30
0
20
40
60
80
100
0
10
20
30

Explain the Second Dimension— Cluster Analysis: K-Means Clustering

Suppose we take the 30 dichotomous responses and conduct kmeans clustering

This provides a simple and quick way to generate groups.

For now, we look at the solution for 3 clusters

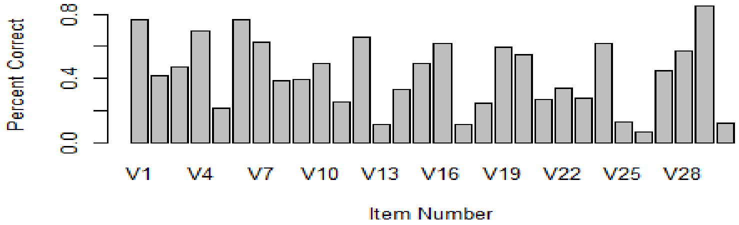
Explain the Second Dimension--Cluster Analysis: K-Means Clustering

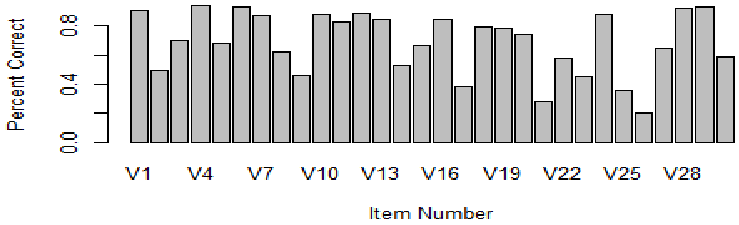
K-means clustering produces three groups. In our data set, these three groups came out as follows:

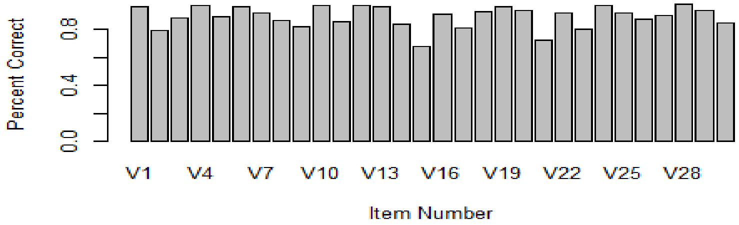
	Mean FCI Score	Cluster Size
Weak	43% 13/30	216
Moderate	69% 21/30	420
Strong	89% 27/30	533

Explain the Second Dimension—
Cluster Analysis: K-Means Clustering

Percentage of correct answers in the three groups.



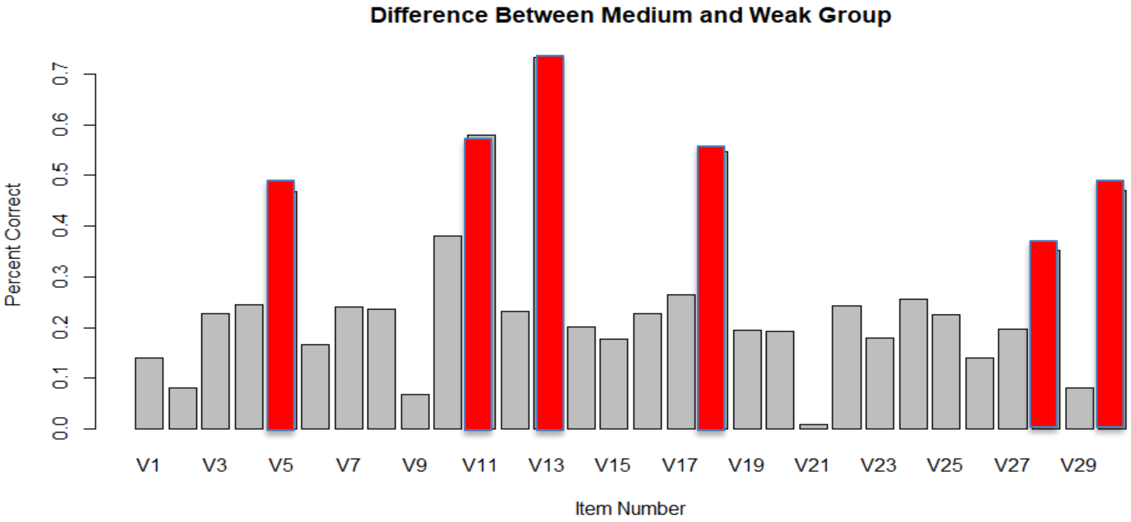




What are the questions on the FCI that most differentiate the groups?

Explain the Second Dimension--Cluster Analysis: K-Means Clustering

To do this we can make a barplot of the differences in the percent correct between the weak and the moderate groups.

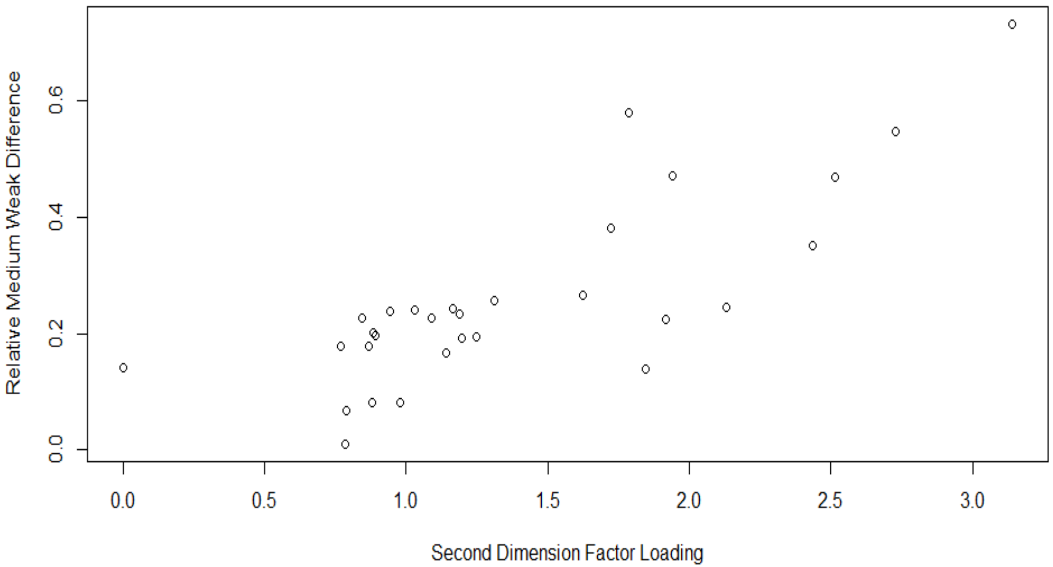


Item 5, 11, 13,18, 28, 30

This information is also implied by IRC graphs.

Explain the Second Dimension--Cluster Analysis: K-Means Clustering

If we plot the relative differences from this histogram against the factor loadings for the second dimension of the IRT model we get the following:

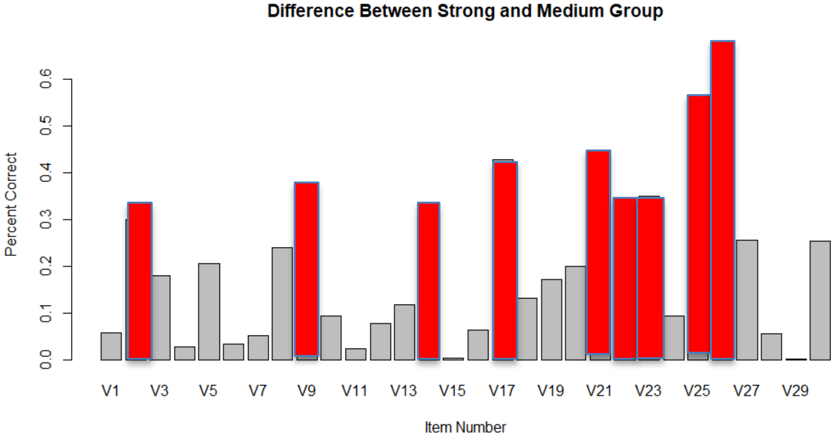


Correlation=0.8

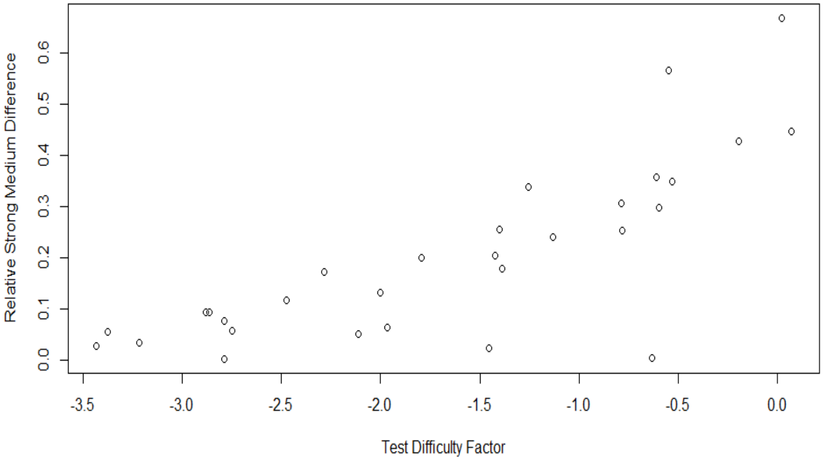
The items on the FCI that show the largest contrast between the low and medium groups are also the items that load on the second factor.

Explain the Second Dimension--Cluster Analysis: K-Means Clustering

Comparing the medium and strongest group. We can ask if the high versus medium contrast corresponds to anything in the IRT analysis. We now highlight a different set of items, and these differences are correlated 0.8 with the test difficulty scores from the 2dim model.



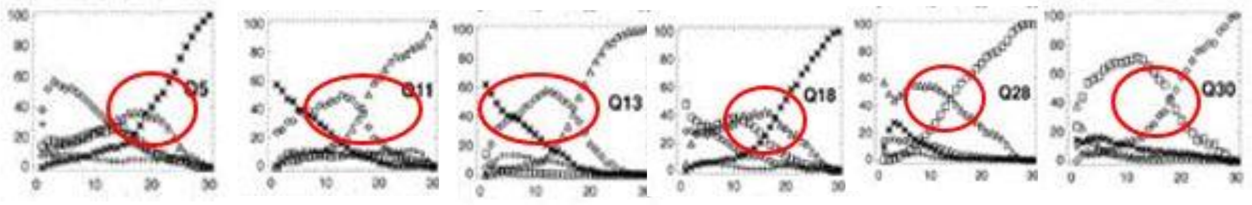
Item 2, 9, 14, 17, 21, 22, 23, 25, 26



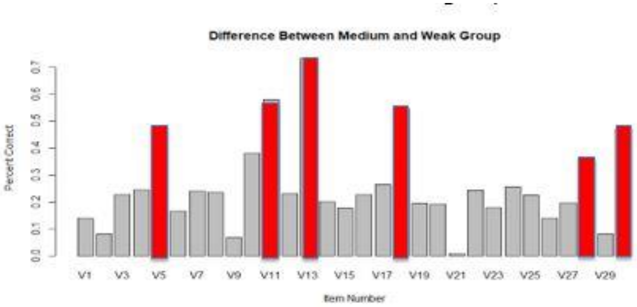
Three different approaches converge

Bayesian IRT: pick up item with higher loadings on the second dimension

Item response curve:



Cluster analysis:



Group	Mean Score
Weak	12.9
Medium	20.7
Strong	26.7

Conclusion

- The current study investigates the dimensionality of the Force Concept Inventory and finds that the two-dimension Item Response Model has a better fit than the unidimensional model.

- We explored the second dimension in two ways:

1. Item Response Curve--the items with high loadings on the second dimension may have strong distractors.

2. K-means clustering-- items with high loadings on the second dimension can differentiate the medium and the weak group.

Future Research

- A clear next step is to use polytomous IRT model.
- Carry out an item analysis with experts
- Investigate evidence of learning progressions, or individual trajectories of improvement.

References

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The physics teacher*, *30*(3), 141-158.

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure?. *The Physics Teacher*, *33*(3), 138-143.

Luo, Y., & Jiao, H. (2017). Using the Stan Program for Bayesian Item Response Theory. *Educational and Psychological Measurement*, 0013164417693666.

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, *80*(9), 825-831.

Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics-Physics Education Research*, *6*(1), 010103.

Scott, T. F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics-Physics Education Research*, *8*(2), 020105.

Semak, M. R., Dietz, R. D., Pearson, R. H., & Willis, C. W. (2017). Examining evolving performance on the Force Concept Inventory using factor analysis. *Physical Review Physics Education Research*, *113*(1), 010103.

Sinharay, S. (2003). Bayesian item fit analysis for dichotomous item response theory models. *ETS Research Report Series*, *2003*(2).

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375-394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British journal of mathematical and statistical psychology*, *59*(2), 429-449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298-321.

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064-1070.

Questions