Item Response Demands, Predicting Item Difficulty, and Validity of Inferences from Test Scores

Steve Ferrara, Jeffrey T. Steedle, and Roger S. Frantz
April 15, 2018

Paper presented in J. Steedle (Chair), *Item Difficulty Modeling: Lessons Learned and Future Directions*, a coordinated session in the annual meeting of the National Council on Measurement in Education, New York

measured progress ACT Questar

Overview

- Goals of the studies
- Literature review
- Results from studies 1, 2, and 3
- Attempt to manipulate item difficulty
- Relevance to item specification, writer training, forms assembly
- In the paper: item-ALD alignment and interpretation validity

Goals of the studies

- Investigate hypothesized content, cognitive, and linguistic item response demands that may predict item difficulty in K-12 achievement tests
- Provide empirical support for significant predictors
- Try to manipulate item difficulty

Item response demands

- Content area, cognitive, and linguistic knowledge and skills required by test items
  - Examinees must recognize, understand, and process to respond to items
- Operationalized as item features
  - E.g., prompts in stem, response options
- May represent proxies for the cognitive processing that examinees activate during testing
  - As indicated in cognitive lab studies (e.g., Ferrara et al., 2004) and IDM studies

Literature review

- 24 studies, 1981-2017
  - More to add
- (a) Reading, literacy, and verbal reasoning; and (b) mathematics and quantitative reasoning
  - Science, insurance certification
- PISA, GRE, state assessments, adult literacy; other
- Mostly OLS; some CART and LLTM

Literature review (cont.)

- R-squares: predicting p-values, IRT locations
  - Reading, etc.: .11-.94, .17-.89
  - Mathematics, etc.: .03-.62, .36-.90
  - Overall: 13 of the 24 studies, GT .50; nine more GT .20

Literature review (cont.)

- Content, Cognitive, and Linguistic demands
  - Also, Item Design, Stimulus demands
- Wide array of content areas, ages and grade levels, testing programs, item response demands ➔
- So far, we find minimal replication of findings on item response demands across these studies
- R-squares indicate lots of practically useful information about item response demands that we can use to manage item difficulty

Example

Table 2. Significant Item Response Demands from the Literature Review

Item Design Demands	
Item/response type, item surface features (e.g., equation in the stem)	Sheehan & Mislevy (1994)
Process variables: literacy (type info requested, type of match); mathematics (operation specificity)	Mosenthal (1998), Kirsch (2001)
Item stems and distractors: reading load, readability, sentence length, negation	Freedle & Kostin (1992), Ferrara et al. (2011), McLeod et al. (2015)
Competing information between passage and distractors, distractor plausibility, content words in correct response and distractors, distractor word frequency/length, vocabulary overlap	Drum et al. (1981), Embretson & Wetzel (1987), Kirsch & Mosenthal (1990), Freedle & Kostin (1992), Mosenthal (1998), Kirsch (2001), Alderson et al. (2009), Lumley et al. (2012), Sano (2016)
Amount of reasoning required to confirm correct response	Gorin & Embretson (2006)
Judgment of item difficulty	Sheehan & Mislevy (1994)

Our three studies

- State and national assessment programs
  - Multiple content areas, schooling levels
- Typical coder training, agreement rates, consensus decision process
- Classification and regression tree (CART) analysis

Results studies 1 and 2

Item Type and Maximum Points demands included

Table 3. Importance Statistics and R-squares for Studies 1 and 2: Empirical Evidence for Item Response Demands

	Study 1				Study 2			
	Lan-guage Arts	Mathe-matics	Science	Social Studies	Grade 4 Social Studies	Grade 7 Social Studies	Grade 5 Science	Grade 8 Science
Item Design Demands[1]								
Item Type[2]	100	100	100	100	97	--	100	--
Maximum Points[2]	NA	NA	NA	NA	100	--	42	--
Conditional Random Forest R-squares (for analyses including all variables in this table)	0.46	0.12	0.21	0.23	0.23	0.09	0.20	0.05

Results studies 1 and 2 (cont.)

Item Type and Maximum Points demands excluded

Table 3. Importance Statistics and R-squares for Studies 1 and 2: Empirical Evidence for Item Response Demands

	Study 1				Study 2			
	Lan-guage Arts	Mathe-matics	Science	Social Studies	Grade 4 Social Studies	Grade 7 Social Studies	Grade 5 Science	Grade 8 Science
Conditional Random Forest R-squares[3] (for content, cognitive, and linguistic demands only)	0.44	0.07	0.08	0.18	0.19	0.07	0.13	0.02

Results study 3

Table 4a. Importance Statistics Study 3: Empirical Evidence for Important Item Response Demands, English Language Arts

	Grade									
	3	4	5	6	7	8 [1]	9	10	11	Mean
Conditional Random Forest R-square	.05	.12	.17	.32	.37	.20	.14	.00	.10	.16

Table 4b. Importance Statistics Study 3: Empirical Evidence for Important Item Response Demands, Mathematics

	Grade/Content Area									
	3	4	5	6	7	8	Algebra 1	Algebra 2	Geom.	Mean
Conditional Random Forest R-square	.40	.50	.39	.33	.44	.47	.47	.33	.36	.37

Significant response demands

Study 1 (high school tests)

- Language Arts, Social Studies
  - Standard/Objective, Indicator, Question Type

Study 2 (performance tasks)

- Grade 4 social studies, grade 5 science
  - Question Type, Relational Complexity, DOK, Linguistic Demands (e.g., Grammatical Density, Vocabulary Density)

Significant response demands (cont.)

Study 3 (grades 3 to high school)

ELA

- Item Design demands, Reading Selection demands including text complexity, Standard/Objective, Command of Textual Evidence (2 grades), Processing Demands (4 grades)

Mathematics

- Item Design demands, Evidence Statement
- No Stimulus or Linguistic demands

Conclusions

- Our studies and literature review indicate many useful item response demands
  - For understanding and becoming more explicit about the KSAs we require of examinees
    - In addition to the stated assessment targets
  - For targeting difficulty levels—and aligning items and ALDs

Conclusions (cont.)

- IDM studies too diverse to find replications of evidence for specific response demands
  - Some exceptions
- K-12 achievement tests in particular
  - Broadly defined constructs, range of item types, developed under challenging time constraints
  - IDM research on an array of content areas, grades, item types, response demands
  - Content area, examinee groups matter
  - Many disappointing R-squares

Conclusions (cont.)

- Manipulating item demands systematically to align with difficulty level targets and achievement levels
  - Not a typical thing for item writers
  - Not typically specified
- A next step

Thank you

ferrara.steve@measuredprogress.org

Jeffrey.Steedle@act.org

rfrantz@questarai.com

measured progress ACT Questar