# Estimating Examinee Intrinsic Difficulty for Providing Greater Specificity of Feedback for Instruction


| Charles Secolsky<br>Venancio L. Fuentes | Bruce Kossar | Eric Magaram |
|---|---|---|
| County College of Morris | Independent Consultant, NJ | SUNY Rockland Community College |
| csecolsky@ccm.edu | bruce@brubumski.com | emagaram@sunyrockland.edu |



*Abstract:* Intrinsic difficulties as opposed to proportion correct scores reflect judgments of what students perceive as difficult. By estimating intrinsic difficulties of items, instructors can gain a greater sense of student misconceptions of content and provide more appropriately focused instruction. Initially, think-aloud protocols established judgment categories (or solution strategies) for twenty basic fraction and decimal problems. A larger scale administration to 238 high school students in general math classes yielded a one-factor solution for the actual correct or incorrect responses to the 20 items and a four-factor solution for the judgments using exploratory principal components factor analyses. BILOG was used to estimate the parameters of the 20 items using a 2PL model while TESTFACT was used to estimate a compensatory multi-dimensional IRT (MIRT) model of the eighty judgments (4 per item) for which only the slopes of the first dimension was plotted along with item response functions of the actual 20 items. Findings indicate that both factor-analytic results and an illustrative example of an IRT and MIRT plot for the misconception of incorrectly cross-multiplying with fraction multiplication identifies specific intrinsic difficulties where instruction can be strengthened. The procedure could have great relevance for students of diverse cultural/ linguistic backgrounds.



*Key Terms*: Intrinsic Difficulty, Student Misconceptions, Developmental Mathematics Instruction


There is and perhaps there always has been a need to improve instruction in many parts of the world. Deficits in student achievement seem to be growing even in light of increases in demand for accountability and in attempts to find solutions to the problems in education. Presently, it appears that the blame in a number of countries including the United States now rests with the lack of adequate instruction. At other times in the not too distant past, poverty, socioeconomic status, race, and the weakness of the family unit have been and continue to be foci of concerns for underachievers. Today, it is mostly the classroom. The present paper represents research that if properly developed and implemented, may be useful in mitigating the problem inside the classroom and outward to the larger community. It is an approach which spends more time on assessing what students know and using that increased knowledge of assessment to aid instruction, particularly where traditional instructional methods have been deemed by existing standards to be failing. This resulting data can support and inspire educators as they rethink and redesign traditional instructional methods.

This proposed approach requires the introduction of a relatively new concept known as *intrinsic difficulty*. Sweller (2010) and elsewhere describes intrinsic difficulty in terms of a theory of cognitive load and it is defined as the germane part of a task that if perceived by the student will probably lead to a correct solution. From a psychometric application of Sweller's theory, the notion of intrinsic item difficulty is proposed for assessing student abilities for potential use by providing more focus and specificity of feedback for enhancing the quality of instruction. Put another way, if we know how students perceive the tasks (in this case test items or knowledge demands), instructors can incorporate this new assessment information into their instruction.

The research spelled out here is based on the premise that the basic mathematics curriculum can change to become more diagnostic. If basic mathematics items can be decomposed with respect to intrinsic difficulty as perceived by students, and if common misconceptions from these perceptions can be identified on a large scale, then curricula can be put into place that may be more useful for instructors who would then be in a better position for remediating students who are experiencing learning difficulties in basic mathematics courses. When piloted with developmental community college students, it was reported that students were more engaged in the assessment. In fact, they appeared to have liked the experience of providing judgments along with providing attempts at solving each of the problems.

While the perceptions of intrinsic difficulty of mathematics items obtained on a large scale may not vary as some would think, recent thought given to this issue suggests that there are apt to be different cultural contexts that would make cultural differences for how data on the assessment of intrinsic difficulties of items are both derived and distributed. (See Henrich, Heine & Norenzayan, 2010 for a discussion of how western culture science may produce experimental psychological results that are different from most other parts of the world.) For purposes of the present investigation, a relatively homogeneous population was used. The population consisted of 238 general high school mathematics students in grades 10-12 from the same high school in the northwest part of the state of New Jersey in the United States.

*Psychometric Foundation of the Problem:*

Item difficulty in classical test theory (CTT) is defined as the proportion of examinees responding correctly to a test item. Item difficulty, the "b" parameter in item response theory (IRT), is estimated using an iterative maximum likelihood method that starts with the proportion correct score. Customarily, both CTT and IRT are not sensitive to the many different ways that examinees find items difficult. Therefore, when test score interpretations are validated, interpretations rely only on some form of correctness or incorrectness based on responses to the actual items. To get around this problem, other evidence is needed in the form of examinee judgments of *intrinsic* difficulty.

Judgments of the *intrinsic* difficulty of test items or the different ways in which examinees find items difficult can lead to more valid test score interpretations than is currently offered by traditional CTT or IRT. For test developers and educators, more could potentially be determined

regarding examinee response processes. Response processes represent one area that is in the realm of validation according to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). As such, this paper is an analysis of the combination of test scores and intrinsic examinee judgments of item difficulty in an attempt to build more sound interpretations of test scores and greater utility for the interpretations (Scriven, 1997).

**Method:** Earlier this year, 22 developmental community college students were administered think-aloud protocols, a method developed by Ericsson and Simon (1993), for 20 different fraction and decimal items. From these tape recorded sessions, the underlying solution strategies of the students were obtained. The transcribed strategies included common student misconceptions on how not to solve these arithmetic problems. The solution strategies were converted into four different intrinsic difficulty categories for each of the 20 items. The interviews with the students took place in a classroom at Rockland Community College during the meeting time of each of the classes. The only instructions students were given were to solve the problems out loud and to continue talking. But, first they received a sample problem from a facilitator who demonstrated how the students were expected to respond. The following example of two questions of what the students were asked to solve in the interview they had, and their thought processes are shown below.

**Find percent notation for 0.372.** Student: Find the percent notation for 0.372. Move the decimal place over 1, 2, 1, 2. I guess I move it over 2 places. That's 37, I think it is just 37%; you get rid of the 2.

**Multiply and simplify $\frac{2}{5}$ x 35.** Student: Multiply and simplify $\frac{2}{5}$ x 35. Uh, you make the 35 a fraction, I think and then you cross multiply. So, no you don't. Oh no, you got to make uh…

Student: Yes, you cross multiply. I have to cross multiplying, 35 x 5, 5x5=25 and 2 up top 5x3=15 and then the 2 up top is 17, so that is 175. I think its 175 and 2 x 175 is. 175 x 2 2x5 is 10, bring the 1 up, 2 x 7 is 14, 15 bring the 5 down, 1 up, 2 x, + 1 is 3 the answer is 350

The responses to the actual open-ended items scored correct or incorrect were collected along with the intrinsic difficulty judgments from 62 community college students. Students attempted to solve the problems and then indicated their intrinsic difficulty judgments in the form of the different solution strategies for each item. Together, both data forms represented 100 data elements for each examinee: 20 actual item scores and 80 intrinsic difficulty judgments. An exploratory factor analysis for the 80 judgments provided considerably greater specificity into the types, composition and interpretation of the resulting factors, especially in the identification of common examinee misconceptions.

If students in high schools become more deficient in basic skills, especially in mathematics, they will create even greater problems for an educated society. Community colleges may have to work even harder to bring student skill levels up to a point where students can become

successful. For this reason, the researchers became involved with a high school and had all students in general math courses in grades 10-12 take the two assessment forms (the actual 20 item test and the 80 judgments form). Two hundred and thirty-eight students took both assessments with the goal of improving the understanding of the internal structure of the responses and ultimately the test score interpretations of the results of the fraction and decimal problems.

**Analyses**: An exploratory principal components factor analysis with a VARIMAX rotation was performed on the actual correct-incorrect responses and a scree plot of the differences in eigenvalues resulted in a one factor solution. The factor solution was then interpreted. Afterwards, IRT ability, difficulty, and discrimination estimates for the regular testing were computed using a 2PL model. An exploratory principal components factor analysis was again performed with a VARIMAX rotation on the judgments (solution strategies) to determine the number of factors on which the judgments loaded and the judgments comprising each factor. The correlation matrix on which the factor analysis was performed consisted of dichotomous data. The judgments turned out to be a multi-dimensional four factor solution. Following, a MIRT analysis of the data was performed to uncover the traits measured by the judgments for each item. However, since the TESTFACT program did not converge after 56 cycles, goodness of fit-statistics were not computed. Figures 1 and 2 show the scree plots for the actual item responses and judgments.

**Figure 1: One-Factor Solution of Item Scores**

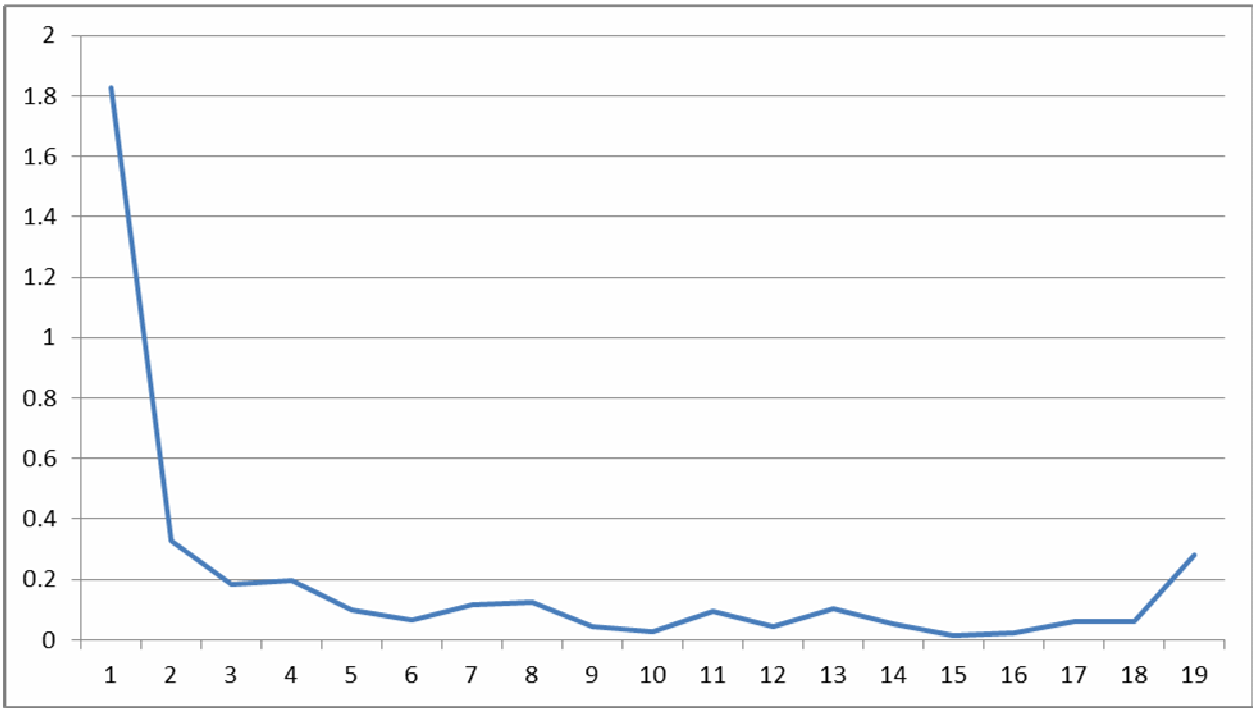
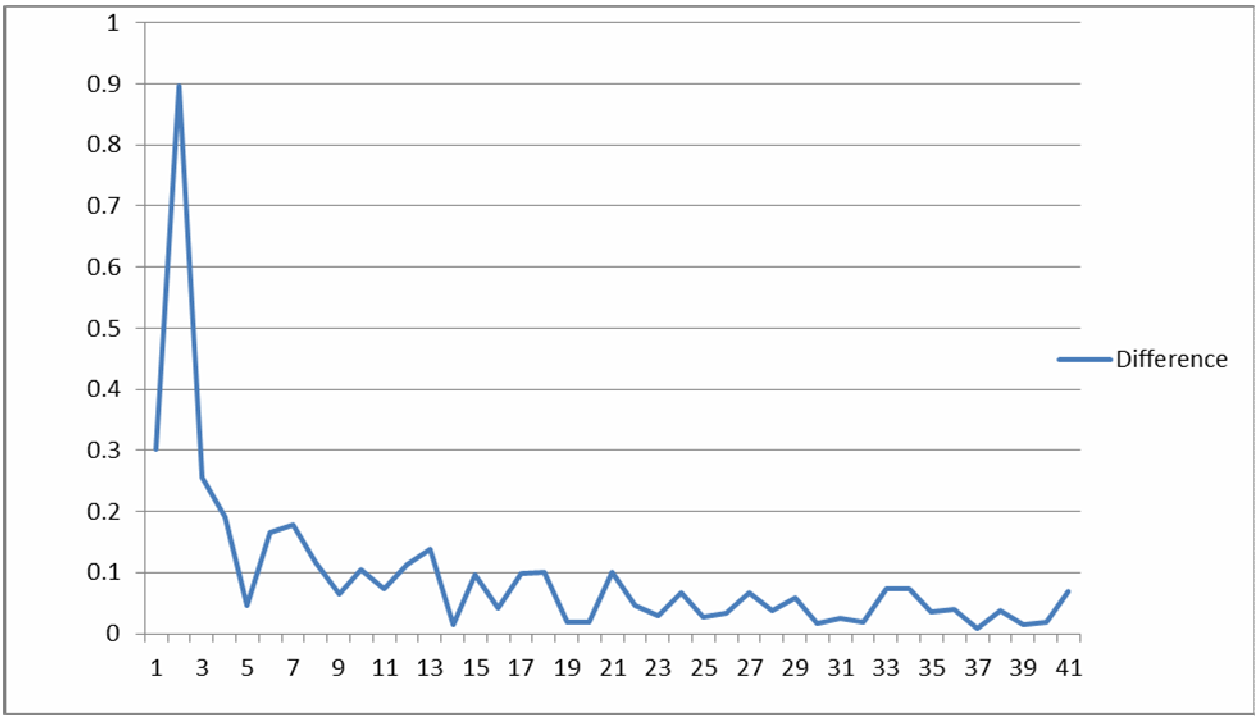


**Figure 2: Four Factor Solution of Judgment Scores**




Each item represented a separate variable and so did each judgment. The relationship between items and their judgments is of critical importance as the goal of the research was to bring a diagnostic focus to what was making each item difficult. The results of the factor analysis while containing the numbers of factors for both the actual test items and the judgments, respectively yielded greater specificity and meaning for the interpretation of the factors and loadings that were obtained.

*Connections of Findings and Research Question:* Item difficulty is defined here in two distinct ways. One way is from the traditional IRT difficulty estimates. Another way of viewing item difficulty is what is difficult from the perspective of the larger modal group of students. The stimuli manifested by the items are the same. But, for the traditional IRT estimates of difficulty, the focus is and previously always has been on getting the answer to the item correct. For examinees, judgments or solution strategies of the difficulty of the items, it is what in each item constitutes the most difficult parts. It is the decomposition of the features of the items that is sorely needed so that the roots of the difficulties with the basic skills mathematics items can be better understood as they exist in the minds of students. The result could potentially provide mathematics instructors with faculty development opportunities for improving upon present diagnostic methods of instruction. It is believed that with this greater understanding by faculty, there will be greater retention rates and ultimately greater success rates for developmental students.

**Results**: Table 1 presents the results of the factor analysis for the actual 20 items using the one-factor solution. Table 2 presents the results for the factor analysis of the 80 judgments using the four-factor solution. The interpretations of these analyses follow the tables.

**Table 1: One- Factor Solution for the 20 actual items.**

| Item | Loading | Actual Item | Key |
|---|---|---|---|
| Q8 | **0.699** | Subtract and simplify. 7/10 – 13/25 | FSS |
| Q10 | **0.629** | Add. 8 1/9 + 7 2/5 | FA |
| Q9 | **0.622** | Add - write as mixed numeral. 6 5/6 + 2 5/6 | FAMn |
| Q11 | **0.615** | Subtract. 9 2/5 – 5 1/3 | FS |
| Q2 | **0.590** | Multiply and simplify. 2/5 * 35 | FMS |
| Q7 | **0.576** | Add and simplify. 7/9 + 5/6 | FAS |
| Q5 | **0.545** | Divide and simplify. 7/4 ÷ 7 | FDS |
| Q6 | **0.533** | Add and simplify. 7/8 + 7/8 | FAS |
| Q12 | 0.433 | Subtract - write as mixed numeral. 27 – 22 ½ | FSMn |
| Q1 | 0.373 | Simplify. 9/15 | FS |
| Q4 | 0.339 | Divide and simplify. 7/2 ÷ 49/4 | FDS |
| Q19 | 0.306 | Find percent notation for 0.372 | Pct |
| Q17 | 0.297 | Find decimal notation. 4/ 15 | Dec |
| Q20 | 0.279 | Find percent notation. 5/8 | Pct |
| Q3 | 0.270 | Multiply and simplify. 3/10 * 43/100 | FMS |
| Q16 | 0.174 | Round to the nearest tenth. 7.8493 | Dec |
| Q18 | 0.151 | Calculate 1/ 4 * 1224 | FM |
| Q14 | 0.100 | Divide - write as mixed numeral. 12 ÷ 1 1/13 | FDMn |
| Q15 | 0.000 | Divide. 7 1/6 ÷ 1 6/7 | FD |
| Q13 | -0.088 | Multiply. 17 4/7 * ¼ | FM |

The items that loaded highest in the one-factor solution had mostly to do with addition and subtraction of fractions, particularly with mixed numbers (Q8-Q12). Other items with relatively high loadings had to do with simplifying. However, the factor solution does not identify what about these problems were intrinsically difficult for examinees and what represented students' misconceptions on how to solve these problems. From the four factor solution of the judgments, it can readily be seen that Wrong choices for solution strategies, particularly for Factors 2 - 4, represent different sets of information that link to the judgments form students completed. For example, the judgments for Factor 2 that loaded the highest were for the Wrong strategy (d) for Q18 and the Wrong strategy (a) for Q3. The strategy for Q18 was "I should have cross multiplied 1224 by 4." The strategy for 3a was "I'm going to cross multiply and get 430 over 300." Clearly, a common misconception was cross multiplying when multiplying fractions across was the Correct solution. Yet, the loadings for items Q18 and Q3 in the one factor solution for the actual items were relatively low, 0.151 for Q18 and 0.270 for Q3, respectively. This indicates at least to some extent that Q18 and Q3 while not well defined as part of the

solution for the actual items, did define Factor 2 and generally extracted more common variance using the judgment data thereby uncovering patterns of examinee intrinsic difficulties.

The three highest factor loadings for Factor 3 of the judgment data were for Q14b, Q7b, and Q8b. The strategies were "First, I find the lowest common denominator by multiplying 9 x 6 =54", "I'm not sure what a mixed numeral is", and "I have to find the lowest common denominator which is 250." Clearly, a common misconception is to multiply denominators to find the lowest common denominator.

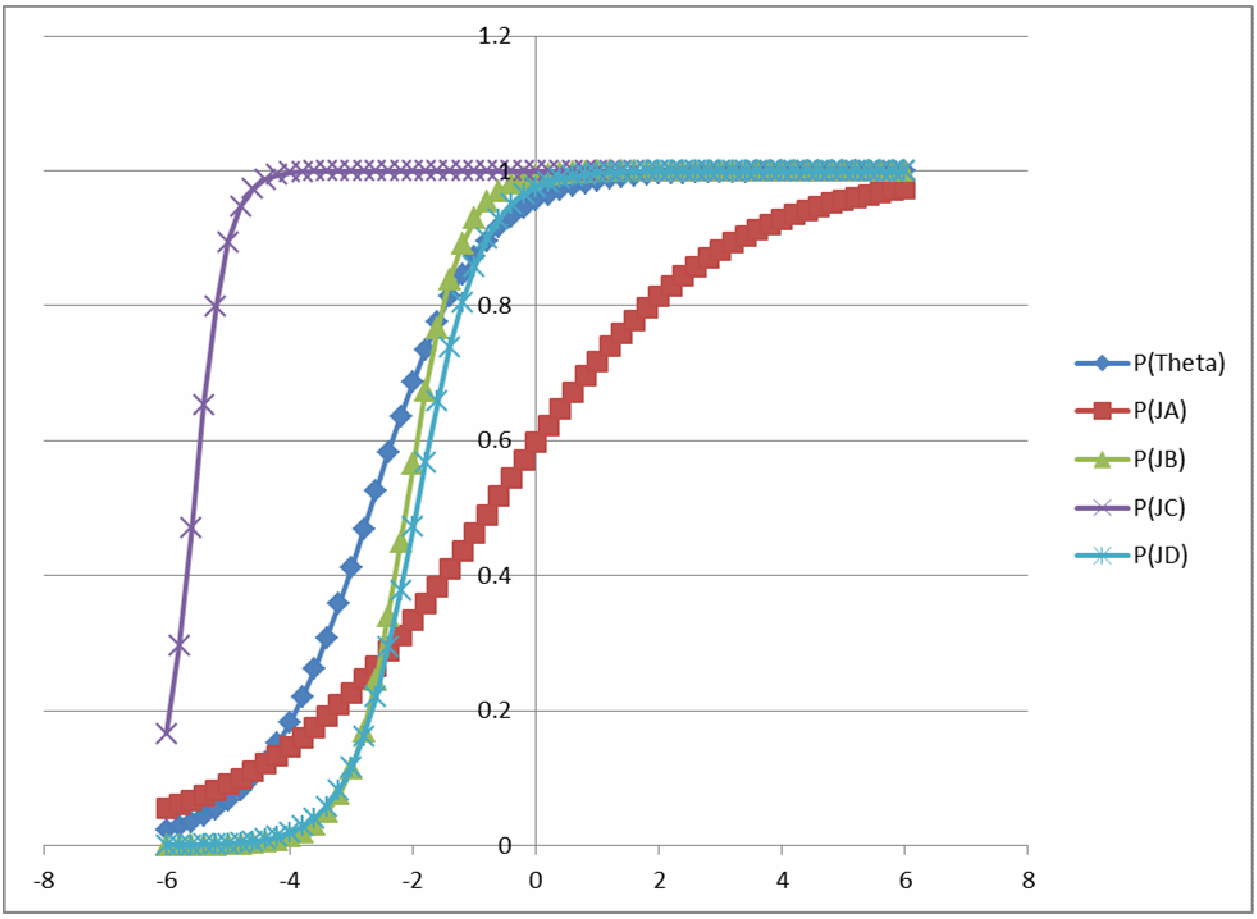
IRFs are produced for Q18 using the 2PL model and for the family of curves comprising the judgment data for this item from the first dimension of compensatory MIRT output. These plots are superimposed for Question 18 for illustrative purposes in Figure 3.

**Table 2: Four-Factor Solution for Judgment Data with Key:**

| Rotated Factor Pattern | Factor1 | Factor2 | Factor3 | Factor4 | New Variables |
|---|---|---|---|---|---|
| C04b-FDS -7 | 0.605 | -0.115 | -0.123 | -0.021 | J4b |
| C05c-FDS -6 | 0.492 | 0.056 | -0.186 | -0.097 | J5c |
| C08a-FSS -4 | 0.480 | -0.217 | -0.186 | 0.391 | J8a |
| W17c-FDec-0 | 0.456 | 0.020 | 0.228 | 0.195 | J17c |
| C07c-FAS -4 | 0.454 | -0.099 | -0.244 | -0.006 | J7c |
| C16c-Dec -4 | 0.438 | 0.132 | -0.077 | 0.146 | J16c |
| C11d-FS -3 | 0.411 | 0.130 | 0.029 | -0.216 | J11d |
| C19b-Pct -3 | 0.410 | 0.057 | 0.026 | 0.185 | J19b |
| C01a-FS -9 | 0.394 | 0.009 | -0.342 | 0.219 | J1a |
| C03d-FMS -4 | 0.392 | -0.218 | 0.030 | -0.152 | J3d |
| C13d-FM -5 | 0.371 | 0.354 | 0.194 | -0.214 | J13d |
| W10c-FA -0 | 0.367 | 0.035 | 0.062 | 0.067 | J10c |
| W12c-FSMn-2 | 0.354 | 0.334 | 0.004 | -0.066 | J12c |
| W20a-Pct -0 | 0.293 | 0.018 | 0.161 | 0.183 | J20a |
| C14c-FDMn-5 | 0.288 | 0.100 | 0.017 | 0.117 | J14c |
| W10b-FA -0 | 0.234 | 0.047 | 0.120 | 0.172 | J10b |
| C09b-FAMn-6 | 0.227 | -0.112 | 0.042 | -0.022 | J9b |
| W06c-FAS -0 | 0.224 | 0.029 | 0.024 | 0.064 | J6c |
| W04c-FDS -7 | -0.257 | 0.108 | 0.107 | 0.213 | J4c |
| W17a-FDec-0 | -0.280 | 0.192 | 0.202 | 0.158 | J17a |
| W19a-Pct -3 | -0.292 | 0.282 | 0.060 | 0.333 | J19a |
| W01d-FS -9 | -0.203 | 0.044 | -0.127 | 0.002 | J1d |
| W18d-FM -4 | 0.052 | 0.572 | 0.037 | -0.068 | J18d |
| W03a-FMS -4 | -0.133 | 0.543 | -0.053 | 0.110 | J3a |
| C15c-FD -3 | 0.079 | 0.494 | -0.034 | 0.056 | J15c |
| W08d-FSS -4 | -0.086 | 0.461 | 0.067 | 0.002 | J8d |
| W06d-FAS -0 | 0.087 | 0.455 | 0.019 | 0.061 | J6d |
| W16a-Dec -4 | -0.083 | 0.449 | -0.021 | 0.082 | J16a |
| W02a-FMS -0 | 0.159 | 0.444 | -0.040 | -0.009 | J2a |
| W07a-FAS -4 | -0.194 | 0.416 | -0.143 | 0.340 | J7a |

| | | | | | |
|---|---|---|---|---|---|
| W20c-Pct -0 | 0.073 | 0.393 | 0.337 | -0.057 | J20c |
| W05b-FDS -6 | -0.180 | 0.384 | 0.233 | 0.034 | J5b |
| W09d-FAMn-6 | 0.109 | 0.328 | 0.102 | -0.083 | J9d |
| W19c-Pct -3 | 0.070 | 0.325 | -0.076 | 0.098 | J19c |
| W10d-FA -0 | 0.171 | 0.245 | 0.105 | -0.122 | J10d |
| W17d-FDec-0 | -0.025 | 0.216 | 0.139 | 0.116 | J17d |
| W08c-FSS -4 | -0.087 | 0.212 | 0.044 | 0.089 | J8c |
| W07d-FAS -4 | -0.031 | -0.220 | 0.122 | 0.080 | J7d |
| W13b-FM -5 | 0.122 | -0.227 | 0.025 | 0.215 | J13b |
| W02c-FMS -0 | 0.186 | -0.258 | -0.072 | 0.055 | J2c |
| W07b-FAS -4 | 0.104 | 0.070 | 0.581 | 0.010 | J7b |
| W14b-FDMn-5 | -0.078 | 0.155 | 0.555 | 0.081 | J14b |
| W08b-FSS -4 | -0.027 | 0.021 | 0.517 | -0.062 | J8b |
| W06b-FAS -0 | -0.090 | -0.065 | 0.437 | -0.019 | J6b |
| W03b-FMS -4 | -0.123 | -0.121 | 0.437 | -0.013 | J3b |
| W19d-Pct -3 | 0.185 | -0.088 | 0.423 | 0.059 | J19d |
| W12a-FSMn-2 | 0.067 | 0.251 | 0.417 | -0.042 | J12a |
| W01b-FS -9 | -0.094 | 0.091 | 0.409 | -0.044 | J1b |
| W13a-FM -5 | 0.122 | -0.042 | 0.407 | 0.227 | J13a |
| W01c-FS -9 | -0.084 | -0.003 | 0.389 | -0.030 | J1c |
| W15b-FD -3 | 0.140 | -0.002 | 0.373 | 0.126 | J15b |
| W11b-FS -3 | 0.047 | 0.211 | 0.324 | -0.064 | J11b |
| W16d-Dec -4 | 0.145 | 0.052 | 0.283 | -0.073 | J16d |
| W04a-FDS -7 | -0.228 | 0.067 | 0.255 | -0.003 | J4a |
| W02b-FMS -0 | -0.188 | -0.140 | 0.255 | 0.161 | J2b |
| W18c-FM -4 | 0.191 | 0.034 | 0.217 | -0.008 | J18c |
| W11c-FS -3 | -0.038 | -0.101 | 0.036 | 0.518 | J11c |
| W10a-FA -0 | -0.235 | 0.217 | -0.016 | 0.444 | J10a |
| W20d-Pct -0 | 0.015 | 0.167 | -0.074 | 0.422 | J20d |
| W05d-FDS -6 | 0.034 | 0.014 | 0.068 | 0.421 | J5d |
| C12b-FSMn-2 | 0.064 | -0.043 | -0.025 | 0.407 | J12b |
| W03c-FMS -4 | 0.076 | -0.181 | -0.097 | 0.392 | J3c |
| W18b-FM -4 | 0.079 | 0.006 | 0.216 | 0.374 | J18b |
| W16b-Dec -4 | -0.030 | -0.151 | 0.276 | 0.373 | J16b |
| W14a-FDMn-5 | 0.139 | 0.155 | -0.225 | 0.370 | J14a |
| W13c-FM -5 | 0.028 | 0.294 | -0.251 | 0.349 | J13c |
| W15a-FD -3 | 0.117 | -0.044 | 0.107 | 0.317 | J15a |
| C18a-FM -4 | 0.208 | -0.031 | 0.008 | 0.300 | J18a |
| W11a-FS -3 | 0.116 | 0.245 | -0.045 | 0.295 | J11a |
| W06a-FAS -0 | -0.043 | 0.144 | 0.086 | 0.290 | J6a |
| W12d-FSMn-2 | 0.065 | 0.010 | -0.053 | 0.278 | J12d |
| C09a-FAMn-6 | 0.050 | 0.116 | -0.064 | 0.268 | J9a |
| W09c-FAMn-6 | -0.080 | -0.068 | 0.136 | 0.265 | J9c |
| W17b-FDec-0 | 0.106 | 0.233 | -0.161 | 0.256 | J17b |
| W15d-FD -3 | 0.241 | 0.069 | -0.018 | 0.244 | J15d |
| W04d-FDS -7 | -0.123 | 0.189 | 0.116 | 0.211 | J4d |

**Figure 3: IRT and MIRT Plots for Question 18 (Q18):**




From Figure 3, the IRF of the actual item P (theta) is plotted in blue. The red IRF curve for judgment A represented the most correct solution strategy for Q18. It stated, "I thought I should change the ¼ to .25 and then multiply by 1224." This judgment was less discriminating over a broader range of ability. The other three IRF judgment curves represented misconceptions, the light blue of which is rightmost and happens to affect slightly more able students than the other two judgment curves. It was judgment D and it was stated as "I should have cross multiplied 1224 by 4." It had the highest loading on the second factor. The green curve represented the judgment B, "By multiplying 1224 by .25, I got 30,600. The purple curve represented the judgment C, "I didn't know I should divide 1224 by 4." Figure 3 demonstrates for the first dimension of the MIRT analysis that the misconception or Wrong judgment that affected the most able students had to do with cross multiplying. This could be stated because it was further to the right on the theta scale.

**Discussion:** Cognitive models for ill-structured content domains are becoming more important as deficiencies and the need for developmental education grow in the United States and around the world. If researchers can begin to understand and model what makes concepts in test items

difficult, then it is incumbent upon us as a research community to be able to diagnose those difficulties. By only examining ability estimates without decomposing items into their difficult parts, we will be standing still and not collecting additional data to understand item difficulty. Since differences were found with the two models, then implications exist for informing instructional methods for teaching such basic arithmetic concepts and computation by providing for faculty development.

By supplementing test scores with judgments of the intrinsic difficulty of items, the responses processes of examinees can be better understood than is usually the case. Including these judgments would make it possible to improve the validation of test score interpretations for two reasons: (1) factor analyses on the judgments of the difficulty of items by examinees provide for greater specificity of examinee misconceptions and (2) MIRT analysis of each item provides feedback on the solution strategies used by examinees for each item and allows for the comparison of MIRT analyses to traditional IRF curves.

**Limitations**:

While possibly decreasing the extent to which error is present in the validation inferences for test scores, there are some drawbacks as well that may come about because of the errors inherent in the solution strategies themselves. More needs to be understood as to that nature of the judgments or solution strategies. It also needs to be determined if instructing students with respect to their major misconceptions about the subjects in question is superior to that of presenting the best solution strategy hoping that it overrides the students' ingrained cognition approaches to each of the test questions.

**References:**

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing:* Washington, DC: Author.

Ericsson, K.A. & Simon, H.A.(1993). *Protocol analysis*: *Verbal reports as data* (revised edition).Cambridge, MA. MIT Press.

Henrich, J., Heine, S.J. & Norenzayan, A. (2010). The weirdest people in the world? *Social Science Research Network.*

Scriven, M. (2002).Assessing six assumptions in assessment. In H.I. Braun, N. Jackson & D.E. Wiley (Eds.: pp. 268-287) The role of constructs in psychological and educational measurement. Mahawah, NJ: Lawrence Erlbaum Associates.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational and Psychological Review, 22*(*2*), 123-38.