Infusing Measurement Theory, Cognitive Science and Instructional Design into Post-Secondary Developmental Mathematics Courses

Charles Secolsky
Center for Instructional Research and Curriculum Evaluation
csecolsky@gmail.com

Stephen H. Levy
Saint Peter's University
drslevy@gmail.com

Developmental students present challenges for post-secondary institutions in terms of readiness for an educated labor force. Underprepared individuals often repeat coursework or drop out, leaving behind a wealth of opportunities for employment. Some models for ameliorating these problems in developmental mathematics propose *easing* of curricular standards. This study however, was aimed at attempting to improve developmental education in mathematics through newly developing theories in psychometrics, cognitive science, and instructional design. Its approach treats-learning deficiency directly. It combines Multi-dimensional IRT, cognitive load theory, and changing lesson delivery, and is validated through randomized group pretest-posttest analysis of covariance using the previous year's final test scores in basic mathematics as the covariate to capture on a large scale what students find intrinsically difficult in mathematics test questions. For n=498 students in the treatment group for whom responses to intrinsic difficulty options were collected and used instructionally, it was found that these students outperformed control group students over the short term of the course. Once mathematics faculty understood and applied this new knowledge of learning difficulties in instruction, they enabled more students to acquire these fundamental skills to-become potentially greater assets to society.

Infusing Measurement Theory, Cognitive Science and Instructional Design into Post-Secondary Developmental Mathematics Courses

The goal of this research was to advance the field of developmental mathematics learning and instruction both methodologically and substantively in the context of postsecondary education. By infusing the latest advances in psychometrics and cognitive science and applying these new developments, particularly in urban areas where many students begin college underprepared to do college-level work in mathematics, this project was a first step at altering the course of developmental mathematics education. It is not a first step at *easing* standards as a number of other postsecondary mathematics initiatives are proposing, but rather it is a direct attack on the present state of affairs via the improvement of learning and instruction.

Recent advances in psychometrics including multi-dimensional item response theory (MIRT), enables researchers to assess at least two abilities simultaneously. In cognitive assessment, a new advance in cognitive science is cognitive load theory (Sweller, 2010), Among other things, it posits that the germaneness or non-germaneness of a task is closely related to the intrinsic difficulty of that task in the minds of students. This difficulty is not the same difficulty defined in classical test theory as the proportion of students responding correctly to an arithmetic test item. In contrast, it is the difficulty defined by the content of the items or task. Using MIRT, the intrinsic difficulties or what is germane to the difficulty of arithmetic test questions can be scored on a large scale to produce sample free multi-trait difficulty parameters in such a way that the resultant information provided for future instruction would contain what is truly difficult about the test questions to some homogeneous population of examinees.

A corollary of this theory is that experts and novices have different ideas about content. Instructionally, it was found that teaching from the traditional expert perspective without considerable attention given to student's prior misconceptions and not altering the instructional delivery accordingly would, with all other things being equal, be inferior to instruction addressing student misconceptions first to undo prior incorrect knowledge structures before teaching from the expert perspective.

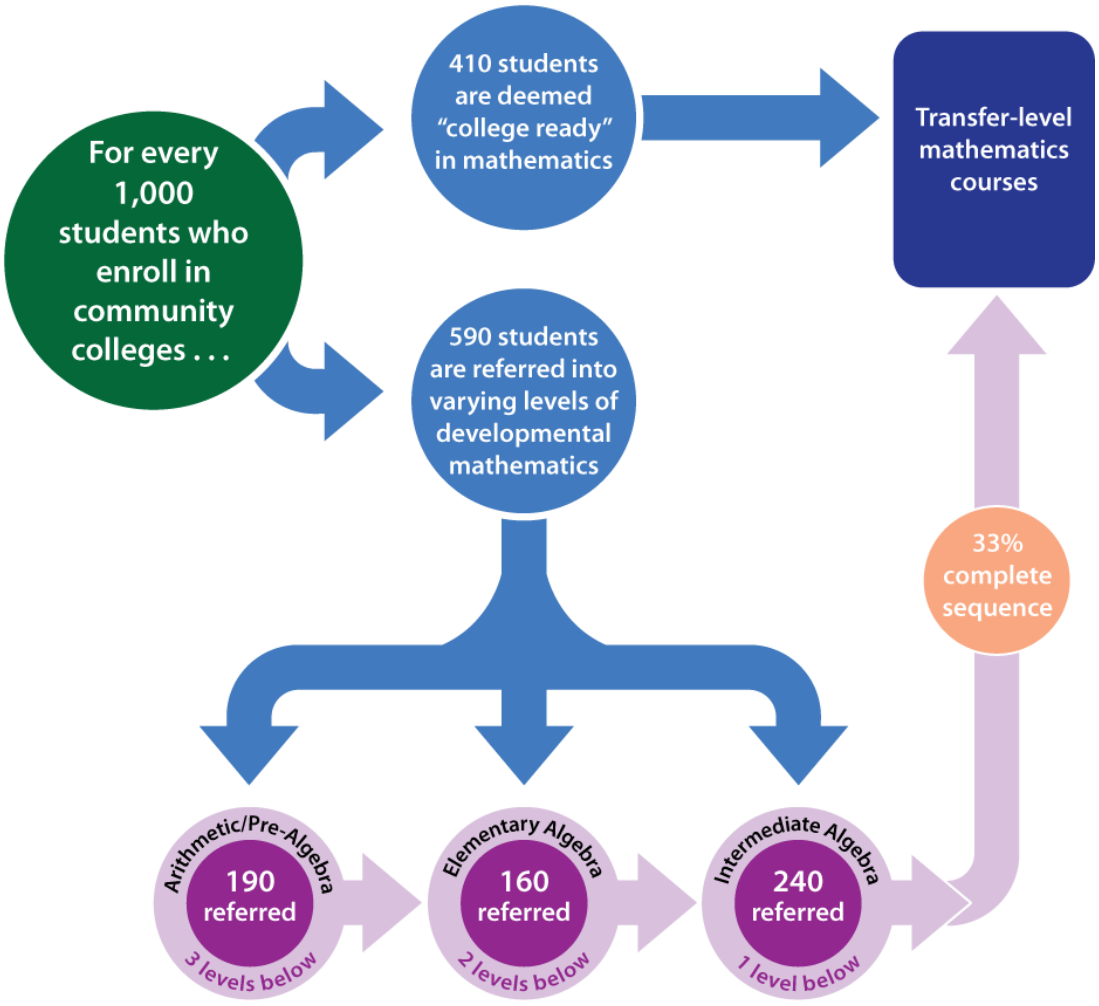
The connections between MIRT, cognitive load theory, and modified lesson delivery have been established by the construction of difficulty options that are produced from qualitative think-aloud protocol data (see Ericsson & Simon, 1993) transcribed from recordings. The congruence between transcriptions and options, as well as the ratings of the technical quality of the options are obtained to improve the validity of claims about item difficulties made from using these options. Once they were understood, the value of the intrinsic nature of difficulties of each item was demonstrated through more traditional quasi-experimental methods, namely analysis of covariance, pretest-posttest randomized control groups design. It was found that the group receiving the difficulty options for review in the course had higher adjusted mean scores on the regular administration of the final exams leading to a greater sense of engagement in mathematics instruction.

Problem Statement: Attempts at improving basic mathematics instruction at the community college level and at four year colleges and universities have not resulted in decreases in the need for developmental courses. Whether students enter college underprepared for college level work even though they have been granted a high school diploma or whether students need

to just brush up on their skills prior to taking an exam, there is a great need in the United States and elsewhere for improving basic skills in mathematics.

There exist different perceptions as to where the blame, if any, lies for the problem of a need for more remedial instruction at community and four year colleges. The college math staffs at community colleges off-handedly say that the high schools do not adequately prepare students. However, it is the high school teachers who are typically trained in instructionalmethods courses. Actually, the problems probably go back to middle school and earlier. Nevertheless, colleges encounter very large numbers of students that are in need developmental education. Nationally, the percentage of students who are referred to one or more developmental courses is about 60% (Attewell, Lavin, Domina & Levey, 2006; Bailey, Jeong & Cho, 2010), and at certain community colleges the rate for need of remediation for entering freshmen is greater than 90% (Kerrigan & Slater, 2010) (see Figure 1).

Figure 1: Typical developmental mathematics sequence at community colleges.



Reprinted with permission from Jenna Cullinane.

The research that is spelled out here is based on the premise that curriculum and instruction in basic skills mathematics can change to become more diagnostic for developmental math courses. Borrowing from the work of cognitive psychologists (see, for example, Sweller, 2010), basic math items can be decomposed with respect to what is germane to the difficulty of these items. Sweller is approaching cognition from the capability of students to provide their own diagnostics for the ultimate purpose of finding out on a large scale what is intrinsically difficult with these items. Such information would be very useful to instructors at all levels so that they can do a better job of remediating students who are experiencing learning difficulties in basic developmental mathematics courses.

The research was predicated on the identification of student misconceptions that are prevalent in basic mathematics. Common student misconceptions were first identified using think-aloud protocols (see Ericsson & Simon, 1993) on the strong possibility that novices and expert practitioners think differently about item content (Ebel, 1956; Secolsky, 1983, Sweller, 1988). From the identified misconceptions in the think-aloud protocols, which were transcribed and converted to item options, MIRT was used for detecting where on the ability scale the different options for an item provided information over the range of ability. Then some instructors were asked to first undo the misconception and only afterward teach from the perspective of an expert. Other instructors representing the control group were asked to teach using traditional methods. It was hypothesized that students in the treatment group where the instructor attempted to first undo the misconception would outperform the control group students taught traditionally. The reason for the difference in means cores was that it was thought that students would be able to identify and engage in the mathematics instruction if they were provided with content prerequisites or co-requisites so that they would be better able to understand the material being taught. These students are likely to be less confused and more attentive to instructors.

Method

Option Development

Based on the work of Ericsson and Simon (1993) students involved in the protocol analysis at Rockland Community College were identified early in the spring 2011 semester from sections of Math 065 – Basic Skills. With the cooperation of two faculty members teaching two different sections of this course, students were asked to go into an adjacent room and provided by researchers with one sample arithmetic question and what is being requested of them in terms of the think-aloud. Students were given five questions one at a time and then asked to think aloud the processes they would use to answer the open-ended math questions. If students ask questions, they would not be provided with any feedback. The only statement the researchers would make after they demonstrate the example is to *please continue talking*. Students were first-year students who are not in the class for repeating the Basic Skills math course. It is the first time they are taking the course.

Once the recordings of the think-alouds were transcribed, the options for the items were developed with assistance of an item writing staff trained to listen to the think-alouds.. The

twenty-five items were then piloted at Hudson County Community College in Jersey City, New Jersey. Ratings of congruence between options and recordings of think-alouds were only after the fact computed similar to methods used by Rovinelli and Hambleton (1977) for aligning items with objectives for classroom instruction. The technical quality of the options was also rated in their own right following the early work of Hambleton (1984). The purpose of the pilot testing was to see if the reasons for why students did not understand the test questions were attractive to the students at this New Jersey community college. From piloting the items on at least three classes at this institution, changes were made accordingly to the options for why the students are having difficulty understanding how to respond correctly to these items.

Large Scale Administration

As previously noted, there were twenty-five developmental mathematics items that would have been developed and pilot tested. A large scale administration of these items took place at the lowest level developmental skills courses using students from class sections at Essex County College, Newark, New Jersey, Rockland Community College in Suffern, New York and Jefferson Township High School in New Jersey .

At each of the participating colleges and high school, instructors of developmental skills courses were identified by the mathematics chair or coordinator. The twenty-five developmental mathematics items were administered at the discretion of the chairpersons. Students were asked to respond to the open-ended item as well as indicate which of the options (misconceptions) students believed would represent a solution strategy that they considered would represent his/her approach to solving the problem. For each item, students were able to choose as many options as apply. Four hundred and ninety-eight students were administered the 25 items. They both answered them and selected as many of the four options per item they deemed were correct strategies. Students in the control group sections in the high school were administered the item options only.

Analyses

An exploratory factor analysis was performed on the responses to the open-ended items and options separately. A scree plot of the items yielded a one factor solution while a scree plot of the option data yielded a four factor solution necessitating the use of MIRT analysis of the data. Goodness of fit-statistics, e.g., chi-square, was used and compared to the IRT and MIRT estimates of ability and difficulty compared to the IRT ability and difficulty estimates for the correct-incorrect scoring of the regular testing using a 2PL model using BILOG-MG. .

Each item represents a variable and so does each option. The relationship between items and their options is of critical importance as the goal of the research was to bring a diagnostic focus to what is making each item difficult. Furthermore, all students in the study would receive a very similar set of items at the end of the semester and it is hypothesized that students in the treatment group would significantly outperform its control group counterparts using analysis of covariance using last year's final exam scores as the covariate.

Connections of Potential Findings with Problem Statement: 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 students' reasons for 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 would potentially provide mathematics instructors with faculty development opportunities for improving upon present diagnostic methods of instruction to be used the same semester for the treatment group. It is believed that the greater understanding that faculty would obtain regarding diagnostic solutions for students with developmental difficulties, the more correct their thinking and the higher the grades students will earn on the final exam for the course.

Cognitive models for content domains are becoming more important as deficiencies and the need for developmental education grow in the United States and throughout the world. If researchers can begin to understand and model what makes concepts in test items difficult, then it is incumbent upon researchers to be able to diagnose those difficulties. By examining ability estimates without decomposing items into their difficult parts, we will not be collecting additional data to understand item difficulty. If differences are found with the two models, then the implication exists for informing instructional methods for teaching such basic arithmetic concepts and computation. If treatment group students significantly outperform control group students at each institution for the same course, then such methods need to be given serious consideration for implementation.

Figure 2 – Samples items with response options

1) Divide and simplify $\frac{7}{4} \div 7$	a) You would have to multiply by ¼. So you get 7/4. 7/4 divided by 7/4 equals 1. b) You start out by changing 7/4 to 1 ¾ and then dividing by 7. c) You should start out by changing to a multiplication problem: 7/4 times 1/7. d) After you have 7/4 times 1/7 you cross multiply to get 49/4 or 12 ¼.
2) Add and simplify. $\frac{7}{9} + \frac{5}{6}$	a) I add the numerators and add the denominators to get 12/15. Then I simplify to get 4/5. b) First, I find the lowest common denominator by multiplying 9 by 6 =54. c) The lowest common denominator is 18. I then multiply 2 by 7 and 3 by 5 = 14+ 15= 29/18 = 1 11/18. d) The lowest common denominator is 36. 28/36 + 30/36 = 58/36 = 1 22/36 = 1 11/18.

Not all of these responses are misconceptions. Some are, some are not.

Formulas

The 3-PL model is planned for the regular administration. If the options produce multidimensional response data from a factor analysis, then a MIRT M3PL model will be used.

3-PL and 2-PL Item Response Functions

$$P \equiv P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}},$$

$$P \equiv P(\theta) = \frac{1}{1 + e^{-1.7a(\theta - b)}},$$

The M3PL model is given by the following.

$$P_i(\underline{\theta}_j) = g_i + (1 - g_i)/(1 + \exp(-1.7(d_i + \underline{s}_i \underline{\theta}_j)))$$

In this model g is the guessing parameter, d is the difficulty parameter, and $\underline{s}$ is a vector of slope parameters, one for each dimension modeled.

Analysis of Covariance (ANCOVA):

$Y_{ij} = \overline{Y} + T_j + b(X_{ij} - \overline{X}) + e_{ij}$

where $Y_{ij} =$ the score of student *i* under treatment *j;*

$\overline{Y}$ = the grand mean on the dependent variable (score on each item);

$T_j =$ the effect of treatment j (receiving options at the onset of the study and final review;

b = common regression coefficient for Y on X;

$\overline{X}_{ij} =$ the score on the covariate for student *i* treatment *j* (Accuplacer placement test score);

X= the grand mean on the covariate;

e_{ij} = the error associated with student *i* under treatment *j*;

First there would be a test for the assumption of homogeneity of regression slopes.

It was later found that the teachers did not teach to the expectations that we had established. The significant result that appears in the following Results section was very possibly due to the fact that the treatment group of students was sensitized or exposed to the test instrument. Therefore, the students were aware of the response options (as in Figure 2) before they were instructed according to our specifications in the new method for addressing the misconceptions in modest detail first. Furthermore, ingrained cognitive structures were still

existed. In other words, their prior misconceptions likely dominated their understanding of the material in this study.. So the improved scores of the treatment group may not be attributable to the new method of addressing misconceptions beforehand to our specified level of detail and therefore may not be valid. Yet, there are redeeming characteristics of the study in the areas of psychometrics and cognitive science, but not in instructional delivery.

Results

It was hypothesized that students who were presented with the *undoing* of a misconception prior to an actual lesson would have greater student achievement. *Undoing* first takes into account obstacles to learning the material while traditional instructional delivery does not. It was thought that teaching in the traditional way would not initially enable remedial students to relate well to the basic math content taught in a regular lesson.

Dependent Variable: Posttes**t** scores
Independent Variable: condition (misconception first=1;traditional only = 2)
Covariates: (Controlling for): pretest; last year's final)

Source	df	Mean Square	F	P
Model	1	27.79	4.51	0.0102
Error	29	6.16		

Dependent Variable: Posttest Scores

Source	df	Mean Square	F	p
Pretest	1	7.79	1.26	0.27
Last year	1	0.23	0.04	0.85
Condition	1	63.25	10.27	0.0033
Error	29	6.16		

This implies that there is a significant difference between the treatment group, which was taught the misconceptions first, and the control group, which was taught the concepts the traditional way. There was less than one-half of one percent probability the result would occur by chance alone (p=0.0033).

Adjusted Mean Scores Out of 20 Items, Controlling for Covariates

Condition	Mean	p
Misconceptions First	8.13	0.0033
Traditional Only	5.01	

Limitations

There is a difficulty in ensuring that the protocols and specifications of the experiment are implemented as they are designed. In this case, it was later found that the teachers of the treatment group introduced the misconceptions (some illustrated in Figure 2) before the detailed instruction began, but they did so in a way too close to the response options in the later posttest and perhaps without the specified level of minimal detail so the students could recall what the misconceptions were. In consequence, confirmation of our results requires further testing more clearly in accord with the design specifications.

Conclusion

The adjusted means indicate that taking into account last year's final exam grade and the pretest score on the assessment form, the group that was taught the misconceptions first averaged approximately 3 points higher on the posttest score ($p=0.0033$).

Evidence has been shown that by attempting to undo the misconceptions first rather than teaching in the traditional way (of first showing a correct method and addressing misconceptions only when students reveal them) results in higher gains. A lesson plan development exercise demonstrating how lessons can be restructured to introduce the misconceptions first provides insight into how to design more effective lesson plans with remedial students in various content domains.

References

Attewell, P., Lavin, D., Domina, T., & Levey, T. (2006). New evidence on college remediation. *Journal of Higher Education, 77*(5), 886–924.

Bailey, T. (2009). *Rethinking developmental education in community college* (CCRC Brief No. 40). New York, NY: Columbia University, Teachers College, Community College Research Center.

Bailey, T., Jeong, D. W., & Cho, S. (2010). Referral, enrollment, and completion in developmental education sequences in community college. *Economics of Education Review, 29*(2), 255-270.

Bhola, D.S., Impara, J.C. & Buckendahl, C.W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(*3)*, 21-29.

Cullinane, J. & Treisman, P.U. (September, 2010) *Improving developmental mathematics education in community colleges: A prospectus and early progress report on the Statway initiative*. National Center for Postsecondary Research.

Ebel, R.L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, 16, 269-282.

Epper, R.M. & DeLott Baker (January, 2009). *Technology solutions for developmental math: An overview of current and emerging practices*. Report funded by the William and Flora Hewlitt Foundation and the Bill and Melinda Gates Foundation.

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

Hambleton, R.K. (1984). Test score validity and standard setting methods. In R.A. Berk (Ed.) Validating the test scores. In R.A. Berk (Ed.) *A guide to criterion-referenced test construction* (pp. 199-230). Baltimore, MD: Johns Hopkins University Press

Kerrigan, M. R.,& Slater, D. (2010). *Collaborating to create change: How El Paso community college improved the readiness of its incoming students through Achieving the Dream* (Culture of Evidence Series, Report No. 4). New York, NY: Columbia University, Teachers College, Community College Research Center.

Maxwell, J.A. (2004).Causal explanation,qualitative research, and scientific inquiry in education. *Educational Researcher*, 33(*2)*,3-11.

Mislevy,R.J. & Riconscente, M.M. (2006) Evidenced-centered assessment design. In S.M. Downing & T.M. Haladyna (Eds.) *Handbook of test development*, (pp.61-90). Mahwah, NJ: Erlbaum.

National Center for Educational Statistics, (2010) US Department of Education. *Integrated Postsecondary Data System,* Fall Enrollment Survey.

Rovinelli, R.J. & Hambleton (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift voor Onderwijsresearch*, 2, 49-60.

Secolsky, C. (1983). Using examinee judgments for detecting invalid items on teacher-made criterion-referenced tests. *Journal of Educational Measurement*, 20(1), 51-63.

Sweller, J. (1988) Cognitive load during problem-solving: Effect on learning. *Cognitive Science 12*, 257-288.

Sweller, J.(2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(*2*), 123-138.