In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.). Proceedings of 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, 1176-1181.

What Were They Thinking? Diagnostic Coding of Conceptual Errors in a Mathematics Learning Software Data Archive

Christine M. Massey (massey@seas.upenn.edu)
Jennifer D. Kregor (kregor@sas.upenn.edu)
Laura E. Cosgrove (lacos@sas.upenn.edu)
University of Pennsylvania, Institute for Research in Cognitive Science
3401 Walnut Street, Suite 400A, Philadelphia, PA 19104-6228

Himchan Lee (himchan@gse.upenn.edu)
University of Pennsylvania, Graduate School of Education
3700 Walnut Street, Philadelphia, PA 19104-6216

Abstract

Decades of research have demonstrated that students face critical conceptual challenges in learning mathematics. As new adaptive learning technologies become ubiquitous in education, they bring opportunities both to facilitate conceptual development in more focused ways and to gather data that may yield new insights into students' learning processes. The present study analyzes data archives from a perceptual learning intervention designed to help students master key concepts related to linear measurement and fractions. Using algorithmic data coding on a database of 78,034 errors from a sample of sixth graders, both conceptual errors and other errors were captured and analyzed for change over time. Results indicate that conceptual errors decreased significantly. This approach suggests additional ways that such datasets can be exploited to better understand how the software impacts different students and how next generations of adaptive software may be designed to code and respond to common error patterns in real time.

Keywords: adaptive learning; conceptual development; educational software; learning technology; mathematics cognition; perceptual learning

Introduction

Prior research on students' conceptual development in mathematics has identified a number of areas in which students persistently make characteristic conceptual errors that are often resistant to standard instruction or procedural practice (National Research Council (NRC), 2001; Vosniadou & Verschaffel, 2004). The present study investigates types of student errors and patterns of change in their performance over time using data archives from sixth graders interacting with an educational software module that was explicitly designed to address such conceptual errors in the domain of units of linear measurement on rulers with both integer and fractional subdivisions. These data were generated as part of a large randomized control trial (RCT) by students in 30 classrooms that were randomly assigned to an intervention condition that used four perceptual learning software modules (Kellman & Massey, 2013) focused on fractions and measurement over the course of their sixth grade year (Kellman, Massey & Son, 2010). Data reported here are from the Linear Measurement Perceptual Learning Module (PLM). Separately reported data from this ongoing study indicate that the PLM intervention demonstrates significant learning gains compared to a control condition (Bowden, Massey & Kregor, 2015; Scull, 2015). For the content of interest in the present analysis, HLM analyses from the RCT indicate a significant treatment effect of the PLM condition, replicated across two cohorts, on a test consisting of multiple choice and open-ended items focused on various aspects of linear measurement drawn from large-scale standardized assessments. The PLM is hypothesized to promote students' conceptual understanding by enabling them to recognize the specific properties of units used to measure continuous extents, to apprehend how whole units and fractional parts of units are represented and enumerated on rulers, and to overcome tendencies to inappropriately apply schemes for counting discrete objects to linear measurement. The present study evaluates this mechanism by examining whether students made the types of conceptual errors that would be anticipated based on the existing research literature on conceptual development for linear measurement and fractions, and, if so, whether the software was effective in helping students overcome known error patterns and move to correct responses.

While this analysis shares some general goals with approaches used in educational data mining and learning analytics (Siemens & Baker, 2010), rather than using machine learning techniques and automated algorithms to discover patterns in responses or to model students and predict responses, the study instead uses algorithmic coding to classify error types predicted from the research literature. Whereas prior cognitive studies of conceptual change in mathematics—particularly microgenetic studies examining learning over time—have generally involved intensive study of relatively small numbers of students interacting with a constrained set of tasks or problems, the current study allows us to examine and code detailed records from 716 sixth graders, each of whom completed an average of 215 open-ended interactive problems over the course of multiple weeks. A total of 157,147 completed problems yielded a pool of 78,034 errors for analysis.

Conceptual Challenges in Linear Measurement

Two areas in which U.S. elementary students perform particularly poorly are fractions and measurement (National

Mathematics Advisory Panel, 2008). Research studies indicate that many students do not recognize that units of linear measurement must have continuous extent, and they instead impose discrete counting schemes on ruler measurement, counting numbered hash marks rather than the intervals between marks (Bragg & Outhred, 2004); Mitchell & Horne, 2008). A familiar result of this misunderstanding is that many students are baffled as to why rulers do not begin at "1." Students also demonstrate consistent errors when measuring with "broken" or partial rulers. Other conceptual difficulties include failing to distinguish between position and distance on a ruler or number line, and not understanding how fractions are represented by subdivisions of units (Ball, 1993; Bright, Behr, Post & Wachsmuth, 1988; Lehrer, Jaslow & Curtis, 2003; NRC, 2001). Also challenging are mapping mixed numbers to rulers and reconciling multiple labels for the same point (e.g., 2/4 and 4/8). Students typically learn a standard computational procedure for converting mixed numbers to improper fractions, but they often lack the ability to flexibly regroup fractions and whole numbers, and, in the context of relating mixed numbers to positions and distances on rulers, the computational procedure may not be productive. The Linear Measurement PLM was specifically designed to address these conceptual difficulties, using a perceptual learning approach in which students directly interact with the targeted structures, relations, and representations across a large and varied set of problems with customized animated feedback on every trial.

Perceptual Learning Software

Perceptual learning (PL) refers to a process by which individuals improve their ability to accurately and fluently extract information coming from the environment in some domain (Gibson, 1969; Kellman & Massey, 2013). This efficient pick-up of information characterizes experts, who selectively attend to relevant features, recognize meaningful patterns, extract higher-order relational structure, and ignore irrelevant variation. Typically, PL occurs through repeated experience discriminating and classifying a wide variety of instances as one engages in a given activity. Recent research (Kellman, Massey & Son, 2010) has demonstrated that principles of perceptual learning can be incorporated into learning software and used to accelerate fluent, expert-like information pick-up in academic symbolic domains such as mathematics and chemistry. Although the term "perceptual" may seem to contrast with conceptual understanding (Kellman & Massey, 2013), in fact, the fluent apprehension of fundamental structures and relationships is often a critical foundation for conceptual understanding. In the present work, PL training is aimed at improving learners' understanding of the structure of whole and fractional measurement units and invariant patterns in how they are represented on rulers and on number lines more generally.

The graphic interface for this PLM consists of an interactive display showing a ball on top of a ruler, as illustrated in Figure 1, which provides examples of a simple integer problem and a more complex fraction problem. Information given at the top of the screen identifies the ball's starting point and then gives either the distance the ball should move and asks the student to indicate the endpoint, or gives the endpoint and asks the student to input the distance it would travel. When the student enters a response, the ball carries out the action, followed by animated feedback indicating whether the response was correct, and, if not, showing how the correct answer compares. On each learning trial, the student sees a unique

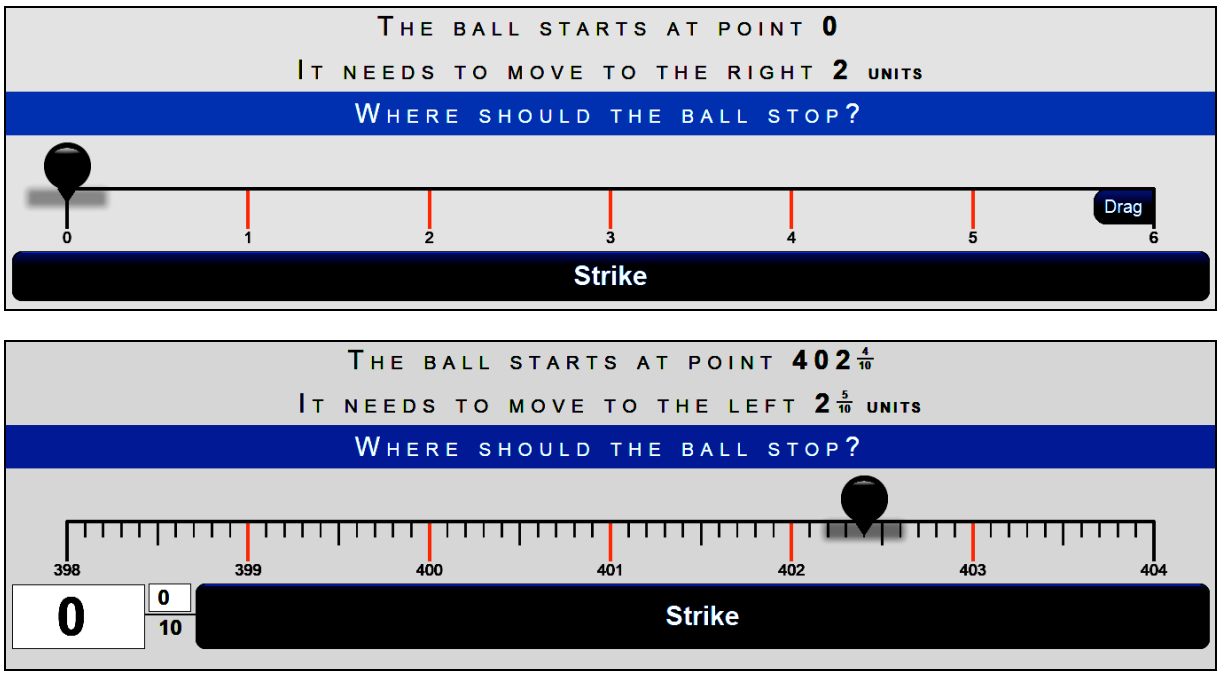

Figure 1: Examples of a simple integer problem (top) and a difficult fraction problem (below).

problem drawn from a very large set of problems organized into eight subtypes, based on whether the problems involve fractions or only integers, whether users enter their responses by moving markers on the ruler or by typing in numerical values, and by whether the unknown in the problem is the final endpoint or the distance traveled on the ruler. Half of the categories consist of simpler integer problems and half are more difficult fraction problems. (Thus one category would be fraction problems on which the distance traveled is given and the user types in the endpoint.) Other problem variations that cut across these 8 categories include whether the ruler is fully or partially labeled; whether the start/end point is 0, 1, or some other point on the ruler (including values in the hundreds); whether the direction of movement is to the right (addition) or to the left (subtraction); and whether the ruler is over-partitioned or congruently partitioned for the units given in the problem.

The software automatically captures time-stamped data, recording every problem seen, the response entered, and the response time (or time-out if no response is entered within 90 seconds). The resulting dataset was analyzed to examine (a) the frequency with which students made the specific conceptual errors anticipated from the research literature on measurement and fractions, (b) what other kinds of errors students made, and (c) whether and how error rates changed as students used the software.

Method

Subjects

Participants in this study were 716 sixth graders in 30 classrooms in schools in a large Northeastern city serving predominantly low-income and minority students. To be included in the analysis, each student had to complete at least 20 problems using the Linear Measurement PLM but did not have to complete the entire PLM. Students used the web-based software during the school day as part of their normal mathematics curriculum.

Procedure

Each unique problem in the software database can be deconstructed into a set of problem parameters. To code errors algorithmically in the large set of student data from participants in this study, we used the problem parameters associated with each problem to create algorithms that operationally define a set of targeted conceptual errors, with a particular focus on miscounting of hash marks and regrouping errors involving fractions. We also developed algorithmic codes for errors related to how students encoded the problems and interacted with the software interface. All codes were built using STATA. Descriptions of the error categories and how they were coded are given below.

Hash Mark Errors Hash Mark errors occur when students focus on discrete hash marks as the unit of measurement and count them starting from one, resulting in answers that are systematically wrong by one. Students can make similar errors with fractional parts of units if they count secondary hash marks on a ruler or number line the same way. When fractions are involved, a student may make the hash mark error only on the integer hash marks, on both the integer hash marks and the fraction hash marks, or only on the fraction hash marks. We designated the first case, along with hash mark errors made on integer problems, as Hash Mark Integer errors, and the latter two cases as Hash Mark Fraction errors.

Regrouping Errors By design, many of the fraction problems in the learning set involve redistributing fractional units from or into integer units across an integer boundary (e.g., the bottom problem in Figure 1.) For right-going Endpoint Unknown problems, this occurs when the sum of the fractional units is greater than "1"; for left-going Endpoint Unknown problems and all Distance Unknown problems, which require subtraction, this occurs when the fraction to be subtracted is greater than the fraction from which it is subtracted. Students often made a characteristic Regrouping error when confronted with boundary-crossing problems. Figure 2 illustrates several examples of such errors on a typical problem. A student would use the correct numerical operation for the integers in the mixed numbers, but use any number of different strategies to deal with the fraction parts: reversing the place of the fractions in order to avoid subtracting the larger from the smaller; answering with either of the given fractions and ignoring the other; or ignoring the fractions entirely. The result for any one of these strategies is nevertheless predictable: an integer answer that is, correctly, the sum or difference of the given integer values but with some incorrect fraction or no fraction appended. This conceptual difficulty is analogous to well-known "buggy algorithms" involving borrowing errors across place value columns in multi-digit arithmetic and in mixed-number subtraction (Brown & Burton, 1978; Fuson, 1990; Scott, 1962). In each case, students fail to process the relational structure of adjacent place values or of fractional and integer units.

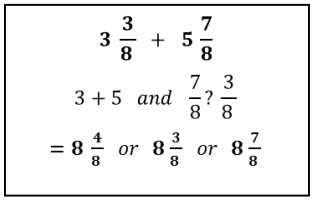

Figure 2: Schematic illustration of possible process for regrouping errors on a problem with a given start point of 3 ⅜ and distance traveled of 5 ⅞.

Problem Encoding Errors In addition to algorithms to capture the targeted conceptual errors described above, algorithms were also developed to capture errors that are specific to the Linear Measurement PLM problem presentation interface. As is often the case in mathematics problem solving, students do not always accurately encode the problem structure in terms of what information is given

and what is to be found. All problems in the PLM involved a triad of a start point, an endpoint, and a distance traveled. The start point was always one of the given values, while endpoint and distance varied between given and unknown roles. One type of problem encoding error occurred when students confused whether distance or endpoint was the unknown. A second type of encoding error occurred when students did not correctly encode the direction of travel (e.g., answered as if the ball moved rightward when the problem specified that it moved to the left).

Given Information Errors Responses were coded as Given Information errors when students entered one of the given values as their answer. Students might do this as a default response when they cannot process the problem structure (similar to "number grabbing" that has been observed when students solve word problems (Bell, Greer, Grimison & Mangan, 1989; NRC, 2001)), or this kind of error may represent a type of disengaged response in which students enter a given value just to enter something.

Unproductive Responses Unproductive Response errors were coded when students pressed "Enter" without giving an answer, timed out without entering any answer, or entered a value that was out of range for the given problem.

Parameter data for each problem were used to create a general code for each error type that would be applicable for all problems or for all problems within a particular subset. For example, a directionality error variable was defined if, for left-going Endpoint Unknown problems, the student's answer for the endpoint is equal to the start point plus the distance. Not all types of errors are applicable for every type of problem, and so the parameter data were used to narrow the test space for particular error codes (e.g. Regrouping errors were only tested on boundary-crossing problems). It is important to note that for nearly all the error codes, the student's answer was flagged only if it corresponded to the answer that would be given if only the named error were made. That is, a student could have concatenated multiple errors—e.g., a Directionality error and a Hash Mark error—and this would not be captured by the error code. Given the risk of miscoding responses unrelated to a complex error combination when operating on a small answer space, however, we chose to avoid concatenating errors.

Results

Out of 78,034 total errors, 38,337 (49.1%) were coded as belonging to a single well-specified error category. An additional 15,753 errors (20.2%) were captured by more than one error code, since the same error could have been made by more than one reasoning process. Since coding of these errors is inherently ambiguous, we removed them for the remainder of the analysis. Errors that were not captured by the algorithmic codes are not considered further in this analysis. Because not all errors can occur on every type of problem, analyses below indicate when the reported frequencies are out of the total of eligible problems rather than all problems.

Table 1 shows the number of students achieving each mastery level. Just over half of the students (52.8%) mastered the entire module, and 60.5% mastered at least 6 of the 8 categories. (To master a category, a student had to complete at least 4 of the most recent 5 problems of that type correctly.)

Table 1: Mastery level by number of students

Mastery Level	N Students	% Students
0	23	3.2%
1	15	2.1%
2	25	3.5%
3	30	4.2%
4	106	14.8%
5	84	11.7%
6	44	6.2%
7	11	1.5%
8	378	52.8%

(Total N of Students = 716)

Table 2 shows the frequency and percentage of each of the captured error types as well as the number of students who made each type of error at least 5 times. As the table indicates, just about half of all errors committed were uniquely captured by the individual codes specified above. Approximately one-fifth of the total errors were the anticipated conceptual errors related to regrouping and to misreading hash marks. Errors captured by multiple codes are not included.

Table 2: Frequency of captured error types

Error	Total Errors Coded	% of Errors All Students	N Students with error 5+ times
Unproductive Responses	11,080	14.2%	479
Given Information	9,588	12.3%	463
Regrouping	8,363	10.7%	464
Hashmark Fraction	5,073	6.5%	382
Problem Encoding	2,533	3.2%	193
Hashmark Integer	1,700	2.2%	111
Total Errors/Total N	78,034		716

Total Problems = 157, 147

To examine changes in the rates of various error types over time, each student's time-ordered sequence of trials was divided into ten phases, from early trials through late trials. Since students completed different numbers of trials, the number of trials falling within each phase is relative to the individual student.

As Figure 3 illustrates, learners typically make steady progress through the PLM, accumulating up to 8 mastery levels as they reach mastery criteria for each category

(typically mastering the easier integer categories first). Figure 3 also shows a distinctive U-shaped curve for average accuracy across time phases. Average accuracy starts at around 68%, as the PLM begins with the simplest integer problems first, and then drops to a low near 50% during the middle of training (phases 5-7), before climbing back up. The steep drop coincides with the appearance of more difficult problems and persists as the easiest problem categories are being retired, which results in students' practice being adaptively focused on more difficult categories. In the last third of training, accuracy again increases as performance improves on harder categories.

Figure 3: Average Mastery Level and Accuracy during Training over Relative Time Phase

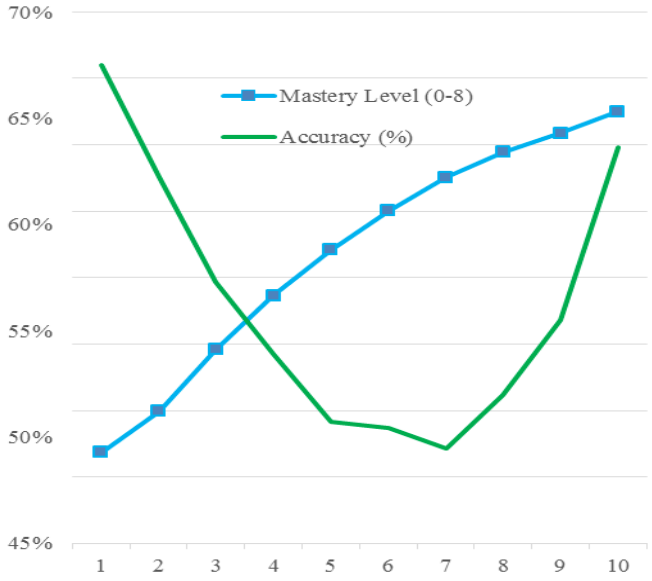

Figure 4: Average Error Rate on Relrelevant Problems over Relative Time Phase

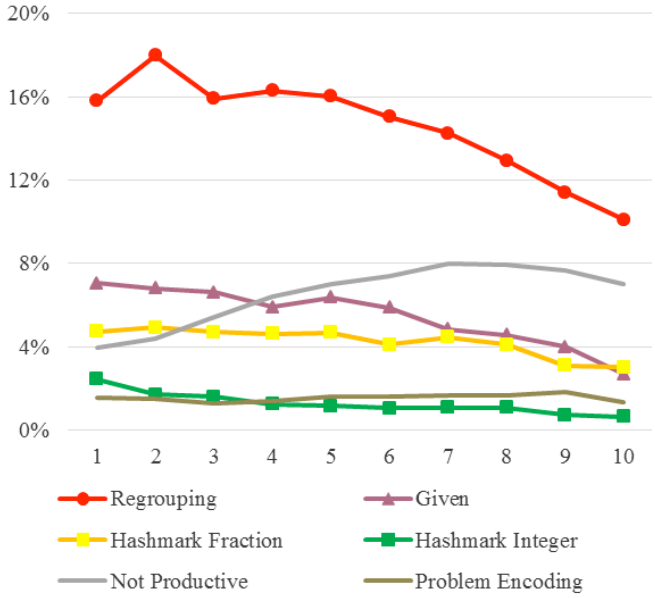

Figure 4 compares the proportion of errors made by each student at each phase of learning, averaged across all students. Regrouping errors showed the highest rate (relative to eligible problems) in all phases, and they decreased steadily in phases 5-10. Notably, the decline in Regrouping errors coincided with an increase in average accuracy and mastery levels across phases 7-10. Both Hash Mark Fraction and Hash Mark Integer errors, which were relatively less common, also decreased across phases. Given Information errors decreased over time, again, with a sharper drop in the later phases. Unproductive Response errors showed a different pattern, initially increasing and then leveling off during the phases in which conceptual errors were declining and correct responses were increasing most rapidly. Problem Encoding errors were relatively uncommon and remained steady across phases.

Repeated Measure ANOVAs were run on each type of error rate to examine mean error rates (averaged across students) across relative time phase. There was a significant effect for nearly all captured errors (using a Huynh-Feldt adjustment for sphericity). Regrouping, Hash Mark Integer, Hash Mark Fraction, and Given Information errors all decreased significantly across phases ($p < .0001$ in all cases). Unproductive Response errors increased significantly across phases ($p < .0001$). Problem Encoding errors did not vary significantly ($p = .07$). Paired t-tests comparing mean error rate at initial and final phases also demonstrated significant results ($p < .0001$) for all error rates except Problem Encoding errors ($p = .62$).

Discussion and Future Directions

The error analyses presented indicate that the Linear Measurement PLM was successful in mitigating several of the specific conceptual errors it was designed to address. Regrouping errors and errors that involved counting hash marks rather than intervals for both integer and fraction rulers declined significantly as students used the software. Most students mastered most or all of the subcategories in the learning set, including fairly difficult fraction problems that required them to be able to flexibly partition and repartition integers and fractions with varying denominators and to add and subtract fractions and mixed numbers. Given that the problems were intentionally varied and required open-ended responses, it is unlikely that students accomplished this formulaically, without gaining some genuine insight into the structure of linear units of measurement and fractions. Indeed, coming to recognize essential structures and relations across novel instances is a hallmark of perceptual learning (Kellman & Massey, 2013).

The targeted conceptual errors showed a distinctive pattern of decline over time, while other types of errors, such as time-outs and unproductive responses, increased during the first half of training. While it may seem paradoxical that some types of errors would increase, this is at least in part a result of the adaptive nature of the software. As students retire certain categories, up until the time all categories are retired, the problems they are seeing come

from not-yet mastered categories and generally become more difficult as students proceed through the module. Thus the pattern of progressive decreases in the targeted conceptual errors indicates that the software was selectively helping learners resolve these conceptual issues.

The methods used in these analyses have significant potential to be extended in ways that would further illuminate students' learning with this kind of adaptive software. Future extensions of this approach will examine the as yet uncaptured errors in the dataset to see if there are more error types that could be coded with well-defined algorithms. Future analyses can also go beyond patterns averaged across students to examine patterns for individual learners or particular subgroups of learners. As the RCT that generated the present dataset proceeds, student-level co-variate data, including demographic data and scores on standardized state tests and on an aligned mathematics test, will become available, which will enable these more detailed explorations. Additional analyses can also investigate at a finer grain how error types interact with specific subtypes or features of problems. While the present analyses have focused particularly on conceptual errors, since that is what the software was primarily designed to address, error data might be analyzed from other points of view. For example, Unproductive Response errors could index disengagement or other motivational or attentional issues for some students. Finally, findings from error analyses of large data archives can be a powerful input to the design process to create new generations of software that are more adaptive in classifying errors in real-time and responding to them in more differentiated ways.

Acknowledgments

We gratefully acknowledge expert assistance from Tim Burke and support from IES, US Department of Education through Grants R305A120288 and R305H06070 to UCLA and the University of Pennsylvania. The opinions expressed are those of the authors and do not represent the views of the US Department of Education.

References

Ball, D. L. (1993). Halves, pieces, and twoths: Constructing and using representational contexts in teaching fractions. In T. P. Carpenter, E. Fennema, & T. A. Ronberg (Eds.). *Rational numbers: An integration of research*. Hillsdale, NJ: Lawrence Erlbaum.

Bell, A., Greer, B., Grimison, L., & Mangan, C. (1989). Children's performance on multiplicative word problems: Elements of a descriptive theory. *Journal for Research in Mathematics Education, 20*(5), 434-449.

Bowden, J., Massey, C. M., Kregor, J. D. (April, 2015). *What predicts successful use and completion of an adaptive mathematics software intervention?* Paper presented at the 2015 Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Bragg, P. & Outhred, L. (2004). A measure of rulers – the importance of units in a measure. *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education, 2*, 159-166.

Bright, G. W., Behr, M. J., Post, T. R. & Wachsmuth, I. (1988). Identifying fractions on number lines. *Journal for Research in Mathematics Education, 19*(3), 215-232.

Brown, J. S. & Burton, R. R (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Fuson, K. (1990). Issues in place-value and multidigit addition and subtraction teaching and learning. *Journal for Research in Mathematics Education, 21*(4), 273-280.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Prentice-Hall.

Kellman, P. J. & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *The Psychology of Learning and Motivation, Vol. 58*, 117-165. New York: Elsevier.

Kellman, P.J., Massey, C.M & Son, J. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, *Special issue on Perceptual Learning, 2(*2), 285-305.

Lehrer, R., Jaslow, L. & Curtis, C. (2003). Developing an understanding of measurement in the elementary grades. In D. H. Clements & G. Bright (Eds.), *Learning and teaching measurement, 2003 yearbook*. Washington, DC: National Council of Teachers of Mathematics.

Mitchell, A. & Horne, M. (2008). Fraction number line tasks and the additivity concept of length measurement. In M. Goos, R. Brown & K. Markar (Eds.). *Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 353-360). MERGA.

National Mathematics Advisory Panel (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council, (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, & B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Scott, L. (1962). Children's concept of scale and the subtraction of fractions. *The Arithmetic Teacher, 9*(3), 115-118.

Scull, J. (April, 2015). *Perceptual learning technology in sixth-grade mathematics education.* Poster presented at the 2015 Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Siemens, G. & Baker, R.S.J.d. (2010). Learning analytics and educational data mining: Toward communication and collaboration. *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.

Vosniadou, S. & Verschaffel, L. (2004). Extending the conceptual change approach to mathematics learning and teaching. *Learning and Instruction, 14*(5), 445-451.