DOCUMENT RESUME

ED 361 349 TM 020 400

AUTHOR Nasser, Ramzi; Carifio, James
TITLE Developing and Validating Sets of Algebra Word Problems.
PUB DATE Apr 93
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Algebra; Analysis of Variance; Construct Validity; Context Effect; *Evaluators; *Interrater Reliability; Mathematics Tests; *Problem Solving; Test Construction; Test Items; Test Reliability; *Test Validity; *Word Problems (Mathematics)
IDENTIFIERS Experts

ABSTRACT

The validation of key contextual features of algebra word problems was studied in two phases. In the first phase, five experts were asked to assess the appropriateness of the concepts in the problems and the adequacy of the assignment of the contextual features to the problems. In the second phase, construct validity was established by having 6 judges rate each of the 16 word problems in random order on the contextual features of familiarity, imageability, and variable type (discrete or continuous). A repeated measures analysis of variance for the construct validity of the key contextual features showed that when one rater or judge and one of the problems were removed, agreement between problems and the criterion were extremely high. When a step-down analysis on each key context feature and variable type was done without one judge, the results indicate a convergence on the constructs devised. In effect, judges agreed with each other, and were correct on 93.5 percent of the ratings, which is strong evidence for both construct validity and reliability of the 16 problems. Seven tables present study findings. (Author/SLD)

"Developing and Validating Sets of Algebra Word Problems"

Ramzi Nasser, UMASS Lowell
James Carifio, UMASS Lowell

ABSTRACT

This paper reports on the validation of propositional relation word problems which was done in two phases. In the first phase, five experts were asked to assess the appropriatness of the concepts in the problems and the adequacy of the assignment of the contextual features (i.e., clothing to the problem's structure) to the problems. In the second phase, construct validity was established by having six judges rate each of the 16 word problems in random order on the contextual features of familiarity, imageability, and variable type (i.e., discrete and continuous).

A repeated measures ANOVA for the construct validity of the key contextual features showed that when removing one rater or judge, and one algebra problem, agreement between problems and criterion were extremely high. Furthermore when doing a step-down analysis on each key context feature and variable type without one judge, the results indicated a convergence on the constructs devised (R=+.95). In effect, judges agreed with each other and were correct on 93.5% of the ratings, which is strong evidence for both construct validity and reliability of the 16 problems.

Paper Presented at the Annual American Educational Research Association Conference in Atlanta, Georgia.

Introduction

Construction and validation of systematic sets of algebra word problems has received little attention from researchers in the area of mathematics education. Further, with the increasing number of factors being identified as effecting students word problem solving processes (e.g., Paige & Simon, 1966; Hinsley et al., 1977; Rosnick & Clement, 1980 and Clement, 1982), better models are needed not only to develop and validate algebra word problems, but also mathematics problems in general than have been used in the past. This study, therefore, will report on the development and validation of key contextual features given to the algebra problem (i.e., the "clothing" of the problem structure), and the data analysis conducted for establishing the construct validity and reliability of the problem set developed, which systematically varies three key contextual features. These key contextual features are: (1) familiarity, (familiar and unfamiliar); (2) imageability, (readily imageable and not readily imageable) and (3) variable type (discrete and continuous).

The effects of familiarity and imageability on the recall of complex material have been widely assessed by several researchers. Normative ratings of familiarity and imagineability by judges have been done for nouns (Stratton, Jacobus & Brinley, 1975; Rubin, 1980) transitive verbs (Klee & Legge, 1976) and adjectives (Berrian, Metzler, Knoll & Clark-Meyers, 1979). Further, key contextual attributes of more complex structures have been studied such as proverbs and sayings (Cunningham, Stanley & Campbell, 1987; Higbee & Millard, 1983). None of these studies, however, employed or attempted to employ any explicit theory to explain or predict the effects of key context features on the process of solving the problems posed. Further, little is said in any of these studies about the construction and validation of the key contextual features alleged be present in the verbal structure of the problems. In a word, the exact nature of the word problems (i.e., stimulus) used in all of these studies actually is a major unverified assumption. It is axiomatic in both research in any scientific field that major assumptions be verified.

Purpose

The purpose of this paper is to report on the validation of a set of 16 algebra word problems. The validation of these algebra word problems was done in two phases. In the first phase, five expert raters were asked to assess the appropriateness of the concepts in the problems and the adequacy of the formulations and contextual features for beginning algebra students. In the second phase, construct validity was established, by having six judges rate each of the 16 word problems in random order on the two contextual features of familiarity, imageability and variable type

(discrete and continuous).

One should note that inter-rater agreement is not synonymous with the reliability of observational measures. For example, all raters may rate the data type in a problem as being continuous when it is discrete. Thus, although their ratings would be reliable, they would not be valid. In addition, construct validity implies that the scores on a test can be meaningfully interpreted in terms of related concepts from a specific theory (Cronbach, 1990). In terms of an information processing model of performance, judges may naively base their rating on their views and theories about the construct rather than the operational definitions and rules specified by the theory being evlauated. Interpretability of scores in terms of the theory and hypotheses being evaluated, therefore, is not something that can simply assumed by researchers based upon (allegedly) face validity considerations.

Algebra Word Problems

Twenty word problems were constructed by the present writers. The set of problems was later reduced to 16. Each of these algebra word problems required students to translate proportions expressed in the word problem into algebraic formulae. The twenty word problems constructed had three different presentation formats modes. These three presentation modes were pictorial, verbal and symbolic representations of the problem. Each mode of presenting a problem had three modes of answering the problem; namely, a pictorial, verbal, or symbolic reponse format. Therefore, students had to process and translate each problem from its presentation mode (pictorial, verbal or symbolic) into a particular reponse format; namely, pictorial, verbal, or symbolic outcomes or answers. Consequently six "cross-translation" combinations (or modes) were possible (see Table 1 for details).

The key contextual features of familiarity, imageability and variable type (discrete and continuous) were the main constructs or interest in these 16 algebra word problems as these attributes individually have been shown to effect performance on arithmetic and algebra problems (e.g., Sims-Knight & Kaput, 1983a & 1983b; Lyda & Franzen, 1945; Sutherland, 1942; Brownell & Stretch, 1931; Washbrone & Osborne, 1926 and Horwitz, 1980). No study, however, has employed algebra word problems that have more than one of these key contextual features let alone all three varied systematically (see Table 1).

When these six modes of cross translation are combined with the key features of familiarity, imageabality, and variable type, one gets the domain of possible algebra word problem type described in Table 1.

As can be seen from Table 1, all of the verbally presented problems were given triads of attributes. The

triads created were: (1) familiar-readily imageable-discrete; (2) familiar-not readily imageable-continuous; (3) unfamiliar-readily imageable-discrete and (4) unfamiliar-not readily imageable-continuous. The verbally presented problems were created to have these features, but these features could not be assigned to the pictorial and symbolic presented problems. A completely crossed rather nested set of features for problems may be constructed by researchers as needed. Hence, the symbolic and pictorial problems in this study were limited to the following attributes unfamiliar-readily imageable-discrete quantities and unfamiliar-not readily imageable-continuous.

Many researchers have viewed student difficulties in the translation and solution of algebra word problems as basically a problem in the student's handling of the verbal structure of the problem (e.g., Mestre, Gerace and Lochhead, 1982; and Mestre and Gerace, 1986). This approach, however, represents a very limited view of algebra word problems and problem solving behavior. No researcher in this area has approached algebra word problem in terms of the various modes in which the problem may be represented (i.e., pictorially, verbally or symbolically), or the mode of representation of the answer to the problem, or in terms of the translation of the relations in these various modes of representations from one mode to another. This very basic limitation in the research literature, therefore, is why we have incorporated modes of representation into our problem set and studies.

Methodology

In the first phase of construct validation, a panel of 5 expert judges were asked to assess the appropriateness of the concepts in the problems and the adequacy of the formulations and contextual features for beginning algebra students. All judges were familiar with the domain of the algebra and had taught algebra, as well as other topics in mathematics, to high school and college students. All of the judges were currently enrolled in a doctoral program in mathematics education and had a Master's degree in either mathematics, engineering, computer science, or mathematics education.

In this first validation phase, the five judges had to read each problem and assess the problems for adequacy, quality and appropriateness in terms of the domain of algebra, as well as in terms students actually being able to do each problem. This phase was an important part of the validation process in order to identify any content or procedural problems that might be present and to ensure that students could read and follow directions given in each problem.

Based on the recommendations of the judges, corrections to the items were made. All of the judges suggestions in terms of wording, coherence, agreement in terminology and symbols, and appropriateness of vocabulary were used to

revise the problems. Problem wordings were simplified, and agreement between symbols and their referents were clarified to eliminate any possible confusion between labels and text.

On the pictorial and symbolic problems, the art work was reviewed and revised to insure the presentation was consistent with the narrative. In addition, drawings were checked with respect to accuracy, scaling and labeling.

The reviewed problems were re-reviewed by two of the judges to ensure that all the suggested changes were made appropriatly. One of the two judges suggested that the verbal problem be changed in terms of its syntax to insure the readability of the problem.

After this final review of the 20 items were made, four items which had the same attributes were dropped from the set which brought the number of items to 16. These sixteen items were administered to 38 high school students taking geometry in inner city high school. Based on the responses students supplied, problems were further reviewed. At this stage, only minor verbal changes were made with respect to two verbal representation. Lastly, one problem was added after being assessed by the five judges.

Rating the Key Contextual Features of the Problem

In the second phase, twelve judges were asked to read each item and rate each as being: 1) familiar or unfamiliar; 2) imageable or unimageable, and 3) discrete or continuous quantities. Each judge, therefore, made 3 ratings for each item. Six raters returned the ratings of the problems out of the twelve selected raters. All the raters were mathematics educators with the exception of one science educator who have had a master's degree in computer science, mathematics, mathematics education and engineering. One math educator held a doctoral degree. All raters where teaching or had taught at high school and college level.

The seventeen problems were given to each rater with a set of instructions and rules based on the adaptation and analysis of the story problem template developed by Kintsch & Greeno (1985). In the rating the key contextual features of problems, judges had to identify the structural components from the propositional statement. For example, given the propositional relation problem:

> "Write an equation, using the variables S and P to represent the following statement: There are six times as many students as professors at this university. Use S for the number of students and P for the number of professors."

four important structural elements were to be identified in the problem. These elements are Noun referents, Qualifiers, Quantities and Relationships.

The noun referents refers to the objects in the problem statement i.e., apples, oranges, professors and so on. Qualifiers function as determiners as well as adjectival

modifiers; e.g., "stupid professors or 2.2 students." Other examples are "speed of a car" or "length increase of a box" which function as determiners of nouns. Thus single noun referents cannot function independently as determiners.

Quantities are the adjective modifiers. They express the cardinality of the objects. For example, 2.2 professors or 6 students, 2.2 and 6 are the quantities. Relationships connect the quantities of the noun referents into a proportion e.g., "there are 20 students for one professor."

A verbal direction prefaces each problem (e.g., write an equation, using the variables S and P). This proposition is verbally stated. In all of the problems the proposition precedes a symbolic, pictorial or verbal representation of the problem. In making the judgements, the raters must relate the propositional statement or direction with the symbolic, pictorial or verbal representation of the problem. For example, in the symbolic presentation "3X=4Y," X and Y will be specifically denoted in the problem by a qualifier or/and noun referent (e.g., X stands for the number of apples and Y stands for the number of oranges). Similarly, pictorial presentations will depict the relationship between the two variables, and also will be specifically denoted by a noun referent and qualifiers in the problem. (All definitions and rules are given in the Appendix A).

Training Raters

Raters were trained by the authors, because raters are are the primary source of potential unreliable observation on each incorrect rating. After the initial ratings and training were made on the problems, all raters were cued or trained with respect to each criterion. Training sessions lasted 5 to 45 minutes depending on the number of implausible reponse each rater made. In the process of training, cross-reference between definitions and rules and implausible responses was made to obtain convergence on the constructs. Because raters would use their own conceptions of the attributes it was necessary for them to converge on the constructs. In most of the cases, there was high convergence as defined in this study. However, on a few occasions some observers did tend to deviate from the criteria and give implausible responses relative to the criteria established for use.

Results and Discussion

Observer agreement is typically assessed by comparing ratings with a criterion. This agreement among raters or judges can be calculated by obtaining interclass correlation coefficients (Haggard, 1958 and Frick & Semmel, 1978). Comparing each judge's ratings with those of a criterion is key to establishing, the construct validity of each item. This analytical approach to establishing the construct validity of items is theoreticaly based and is derived from a logical analysis of the analytical problem inherent in the

data collected. According to Frick & Semmel (1978), assessing the degree of agreement between raters and criterion is the more appropriate analysis than interjudge agreement when decisions are being made about individual judges and criteria.

Given the above points, three ratings were made on each problem by six raters. Each raters, therefore made 51 judgements yielding 306 (6x51) judgements in all. Analysis of these data should establish construct validity by comparing all judges with all criteria. This later criterion satisfies the rules of correspondence which connect the theory and data and examine whether or not the data satisfy the theory (Haggard, 1958).

To assess rater judgements of the presence or absence of the key contextual features in each algebra problem (i.e., construct validity) two important statistics were used. First, an F statistic which was used to assess the question of difference between raters across the 51 judgements made. Second, the inter-class reliability, R is used to explain the question of similarity of profile scores on certain measure with respect to a criterion. The coefficient R is also related to the interclass or product moment coefficient of correlation and is used as a index for the similarity of sets of scores. Hence, R determines the consistency within classes and is thus an estimate of reliability.

To calculate these statistics, a repeated analysis of variance was performed on all the items, followed by a seperate step-down analysis on all items with respect to each classification of its key context features (i.e., familiarity imageablity and variable type) in order to obtain an estimate statistic of differences among raters and among items.

Table 2 presents the results of the 6x51 one-way repeated measures analysis of variance done for the 6 raters and 51 attributes judged. As can be seen from Table 2, significant main effects was found among the items ($F=(50,250)=2.2, p<.05$). These results show that judges rated the attributes to some problems in a different manner.

To assess which features were rated differently by which judges or which items, a step-down analysis of each feature for each item was done. Tables 3 presents the results of a one-way repeated ANOVA (6x17) for the 6 raters and the 17 problems for the key feature of familiarity. No significant difference or interaction were found between raters or items, for the key featrue of familiarity.

Table 4 presents, a one-way repeated measure ANOVA (6x17) done on the ratings for the key feature of imageablity. A high significant difference ($F(5,80)=2.6$, $p<.05$) was found among the raters and items (see Table 4) for the readily imageable and not readily imageable attributes. Similarly, Table 5, is a one-way repeated measure ANOVA (6x17) done on the ratings of the key feature of variable

type. No significant differences were found between raters on items.

To Examine the pattern of judgements further, the "correct" feature ratings were summed over all judgements on all problems and divided over the total number of ratings to produce a "correct response index," for each judge. As can be seen from Table 6, Judge 3 had the lowest correct index score several of imageable items (i.e., readily imageable and not readily imageable.) Removing this judge from the analysis and again performing a repeated measures ANOVA, the results (see Table 7) showed a significant main effect on the items, but no significant differences were found between raters. By further analyzing the disagreement within items, the data showed the highest disagreement among raters on a pictorially presented problem: namely, an item which did not have having a readily imageable attribute. When a repeated ANOVA (5x16) was performed between raters removing this pictorial item, as well as judge 3 from the analysis, no significant main effects was found among raters ($F(4,60)=1.0$, $p>.05$) and items ($F(15,60)=1.0$, $p>.05$). These results indicate that most of the difference observed in the data could be assigned to rater 3, who could have misread the directions or did not maintain her or his skills when trained.

Overall, the repeated measures ANOVA results were very positive. Only 6.5 % of the ratings of the key context features were incorrect. The raters, therefore, agreed with each other and were correct on 93.5% of the ratings, which is both strong validity and strong reliability evidence for these 17 algebra problems. Further, there were 8 problems, where all six raters agreed on all of the key contextual features and were correct in their observations, while in 9 problems, there were only three or less raters who disagreed on the key contextual features present in the problem. When one item which had a high incorrect response rate was removed from the analysis, no statistically significant differences was found among raters across all items which is extremlely strong evidence for the construct validity for the 17 of these items.

Reliability of the Ratings

Important sources of variation in the validation of problem sets arises with differences among judges and/or items. Thus, when two or more sources of variation or error are present in the measurement, the pearson product moment correlation, as an estimate of reliability is used but this measure fails to provide an unbiased measure of reliability (Haggard, 1958). In this view, the intraclass correlation, R provides an estimate of reliability and computed by the analysis of variance technique. Consequently, R describes to the relative degree of consistency among sets of rater profiles or scores.

The reliability for the similarities of profiles in the Haggard (1958) approach is based on the degree of similarity of rater scores profile. Because the agreement obtained is based on a right-wrong comparison of the observer with the scoring criteria, the calcualtion of R differs from its typical calculation method. Thus, the similarities between profiles of responses on the scored attributes is obtained by an inversion of the conventional formula; $R=1-R_{ck}$, were R_{ck} is the measure of dissimilarity among the profiles or the judge item interaction. Therefore, when the judges ratings are identical to the criteria, the interaclass correlation R score will literally be equal to 1.

The interrater reliability score among the 51 ratings for all the attributes was at R=+.95 for the six raters, using Haggard's (1958) ANOVA procedures in computing the interclass R. The interrater reliability for familiarity, imageability and variable type classifications were at R=+.93, +.95 and +.97 respectively. The lowest interrater reliability was observed on the familiarity classification as was expected because of a scattered ratings in the profile between different raters within different items. As can be seen from these coefficients, both correctness and agreement level were extremly high.

The results obtained from the two sets of analysis done indicate that the interjudge reliabilities and construct validity of the 17 algebra word problems we devised were extremly good. Raters could objectively and reliably discriminate within each of the three categories of key context features, the correct attributes present in each problem after training. Various analysis showed, however, that one pictorially presented problems which had unfamiliar and readily imageable features had the highest level of disagreement among raters. This particular finding would seem to contradict the some what naive but popular view that anything presented pictorially is depictable or easily imaged (Kosslyn, 1983). Therefore, when multiple key contextual features are present in an algebra word problem, it seems from the evidence currently available that features may interact and interfer in the processing of the problems representation and solution both cognitively and/or unconsciously. Data from students actual performance on these 17 word problems dispute the information processing model. General problem-solving performance was lower on those problems with the familiar-readily imageable features than those with the unfamiliar-not readily imagebale features. These results dispute the information processing model as the key context features were assigned to the problem to reflect on the theoretical basis of the model

Another possible view of this finding is that judge 3, who had the highest number of incorrect responses, may have given the ratings long after the training session which

states that the agreements following initial training was found to decrease as a function of time. Eliminating judge 3 and the pictorial presented problem showed no significant difference on the profiles of the agreement among main effects of raters and items.

Conclusion

We sought to develop and validate a domain referenced set of algebra word problems that systematically reflected key factors (or contextual features) identified from the research, which influence students' abilities to solve such problems. No study of students abilities to solve algebra word problems has employed problems that have more than one of the key contextual features identified (i.e., familiarity, inageability and variable type) from the literature. The 17 algebra word problems systematically varied several key context features the validity of these systematic variation of key features was confirmed empirically by six judges. The domain referenced set of 17 algebra word problems we developed will allow mathematics educators to better assess and qualitatively evaluate the effects of current algebra course and texts. And to conduct better and more sophisticated research in this important area as well as to make prior studies in this area more interpretable in terms of their results.

The algebra word problem development and validation model we used consisted of two phases. One, a consensual assessment phase and a second construct validity phase. This model was relatively easy to implement and carry out and produced high quality items of the kind being proposed by the authentic assessment movement in all areas of education.

Analysis of the data gathered revealed the following about construct validity and reliability of our 17 algerba word problems. First, the items had a high degree of consistency among sets of intraclass scores. Although the F statistic showed significant difference between raters, this difference was due to one item and one judge. When the judge and the item were removed from the analysis this difference was removed. The raters, therefore, agreed with each other and were correct on 93.5% of the ratings, which is both strong evidence for construct validity and strong reliability for the 17 algebra problems.

We recommend that the item development procedures described in this study be generalizable to other large scale assessments of item sets with a broader mathematics domain. Also we recommend that the validity of the key context features of our algebra word problems be assessed in a follow by a group of students characteristics of those who will attempt these problems. Such a follow up study would allow comparisons to be made between novice and expert raters which could be highly informative on a number of theoretical issues and outstanding instructional questions.

Table 1

A Descriptive and Conceptual Characterization of the Domain of Algebra Word Problem.

Mode of Representation and Cross Translation	Key Contextual Features FI/D	UI/D	FU/C	UU/C
Verbal to Symbolic	1	1	1	1
Symbolic to Verbal		1		1
Pictoral to Symbolic		1		1
Symbolic to Pictoral		1		1
Verbal to Pictoral	1	1	1	1
Pictoral to Verbal		1		1

FI/D= familiar-readily imageable-discrete
UI/D= unfamiliar-readily imageable-discrete
FU/C= familiar-not readily imageable-continuous
UU/C= unfamiliar-not readily imageable-continuous

Table 2

Repeated Measures ANOVA on Ratings of all Assignments of the Seventeen Translation problems (N=6)

Source	df	Mean Squared	F	p
Raters	5	.07	1.4	> .05
Items	50	.11	2.2	< .05
Error (interaction)	250	.05		

Table 3

Repeated Measures ANOVA on Ratings on Familiar; Unfamiliar Assignments of the Translation problems (N=6)

Source	df	Mean Squared	F	p
Raters	5	.11	1.375	> .05
Items	16	.15	1.875	> .05
Error (interaction)	80	.08		

Table 4

Repeated Measures ANOVA on Ratings of Readily Imageable and Not Readily Imageable Assignments of the Translation problems (N=6)

Source	df	Mean Squared	F	p
Raters	5	.13	2.6	< .05
Items	16	.13	2.6	< .05
Error (interaction)	80	.05		

Table 5

Repeated Measures ANOVA on Ratings of Discrete and Continuous Quantities Assignments of the Translation problems (N=6)

Source	df	Mean Squared	F	p
Raters	5	.02	.67	> .05
Items	16	.05	1.67	> .05
Error (interaction)	80	.03		

Table 6

Summation Index of the Correct Responses for the Readily Imageable and Not Readily Imageable Attribute

Judge	Readily Imageable	Not Readily Imageable
1	1.00	.88
2	.89	1.00
3	.67	.88
4	1.00	.88
5	1.00	1.00
6	1.00	1.00

Table 7

Repeated Measures ANOVA on Ratings of Readily Imageable and Not Readily Imageable Assignments of the Translation Problems by Removing one Rater (N=5)

Source	df	Mean Squared	F	p
Raters	4	.02	.67	> .05
Items	16	.06	2.00	< .05
Error (interaction)	64	.03		

Reference

Berrian, R., Metsler, D., Knoll, N. & Clark-Meyers, G. (1979). Estimates of imagery, ease of definition, and animateness for 328 adjectives. Journal of Experimental Psychology: Human Learning and Memory, 5, 435-447.

Brownell, W. and Stretch, L. (1931). The effect of unfamiliar settings on problem-solving. Duke University Research studies in Education, Durham, N.C.: Duke University.

Clarkson, S. (1978). A study of the relationship among translation skills and problem-solving abilities. Unpublished Doctoral Dissertation. University of Georgia, Georgia.

Clement, J. (1982). Algebra word problem solutions: thought processes underlying a common misconception. Journal For Research in Mathematics Education, 13(1), 16-30.

Cronbach, L. (1990). Essentials of Psychological Testing. Fifth edition. New York: Harper & Row

Cunningham, D., Stanley, R. & Campbell, A. (1987). Relationship between proverb familiarity and proverb interpretation: implication for clinical practice. Psychological Reports 60, 895-898.

Frick, T. & Semmel, M. (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48(1), 157-184.

Haggard, E. (1958). Intraclass correlation and the analysis of variance. New York: Dryden Press Inc.

Higbee, K. & Millard, R. (1983). Visual imagery and familiarity ratings for 203 sayings. American Journal of psychology, 96(2), 211-222.

Hinsley, D., Hayes, J. & Simon, H. (1977). From words to equations: meaning and representation in algebra word problems. In Marcel Just & Patricia Carpenter (Eds.) Cognitive Processes in Comprehension. Hillsdale, N.J: Lawrence Erlbaum Associates.

Horwitz, L. (1981). Visualization and arithmetic problem solving. Los Angeles, CA: Paper presented at the annual meeting of the American Educational Research Association.

(ERIC Document Reproduction Service No. ED 202695).

Kintsch, W. & Greeno, J. (1985). Understanding and solving word arithmetic problems. Psychological Review, 92(1), 109-129.

Klee, H. & Legge, D. (1976). Estimates of concretness and other indices for 200 transitive verbs. Journal of Experimental Psychology: Human Learning and Memory, 2(4), 497-507.

Kosslyn, S. (1983). Ghosts in the mind's machine: creating and using images in the brain. New York: W. W. Norton & Company.

Lyda, W. & Franzen, C. (1945). A study of grade placement of socially significant arithmetic problems in the high school curriculum. Journal of Educational Research. 39(4), 292-295.

Mestre, J. P. & Gerace, W. J. (1986). The interplay of linguistic factors in mathematical tasks. Focus On Learning in Mathematics, 8(1), 58-72.

Mestre, J. P., Gerace, W. J. & Lochhead, J. (1982). The interdependance of language and translational math skills among Bilingual Hispanic engineering students. Journal of Research in Science Teaching, 19(5), 399-410.

Paige, J. and Simon, H. (1966). Cognitive processes in solving algebra problems. In Kleinmuntz, B. (Ed.), Problem solving: research, method and theory. New York: John Wiley & Sons, Inc.

Rubin, D. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. Journal of Verbal Learning and Verbal Behavior. 19, 736-755.

Rosnick, P. & Clement, J. (1980). Learning without understanding: the effect of tutoring strategies on algebra misconceptions. Journal of Mathematical Behavior, 3, 3-27.

Sims-Knight, J. and Kaput, J. (1983a). Misconception of algebraic symbols: representation and component process. Proceeding of the International Seminar. Ithaca, NY: State University of New York and Cornell University. (ERIC Reproduction Service No. ED 242 553)

Sims-Knight, J. and Kaput, J. (1983b). Exploring difficulties in transforming between naturaql language and image based representations and abstract symbol systems of mathematics. In Rogers, D. & Sloboda J. (Eds). The aquisition of symbolic skills. New York: Plenum press.

Stratton, R., Jacobus, K., Brinley, B. (1975). Age-of-aquisition, imagery, familiarity and meaningfulness norms for 543 words. Behavior Research Methods & Instrumentation, 7, 1-6.

Sutherland, J. (1942). An investigation into some aspects of problem solving in arithmetic. British Journal of Educational Psychology. 12, 35-46.

Washborne, C. & Osborne, R. (1926). Solving Arithmetic problems. Elementary School Journal. 27, 219-226.