You Get What You Assess: Evaluating Achievement in Mathematics

Using a Constructed Response Test Formatted Model

Kathy A. Gullie

University at Albany/SUNY

For additional information contact:

Kathy A. Gullie, Ph.D.

Evaluation Consortium

Dutch Quad, Beverwyck B7

University at Albany/SUNY

Albany, New York 12222

Phone: 518 442-5027

Email: kgullie@albany.edu

Paper presented at the Annual Meeting of the American Evaluation Association, Washington DC, October, 2013.

You Get What You Assess: Evaluating Achievement in Mathematics

Using a Constructed Response Test Formatted Model

Kathy Gullie

University at Albany/SUNY

ABSTRACT

This paper examines the utility of constructed/extended response questions as a means to promote conceptual understanding of mathematics and facilitation of positive student outcomes. Currently, investigations of item formatting related to facilitating student understanding and achievement in mathematics is limited. Because items that prompt students to construct a response are thought to support conceptual development, we argue that such strategies will enhance knowledge transfer, and in the process, develop students’ flexibility in thinking through multiple contexts. Discussion in the paper centers on research supported by the impact of testing on classroom instruction, with special emphasis on how such an impact can be evaluated and used to enhance conceptual learning while eliciting positive outcomes from all formats of testing items. This study uses a multivariate correlational analysis of longitudinal archival test results (grades 3-5) from an urban upstate New York elementary school to describe relationships between open-ended/constructed response/short answer items on the State mandated Mathematics Tests and student proficiency.

You Get What You Assess: Evaluating Achievement in Mathematics

Using a Constructed Response Test Formatted Model

Kathy Gullie

University at Albany/SUNY

*Purpose of the paper*

This evaluation focused on student responses to constructed/extended response questions on a statewide test. A multivariate correlational design examined a priori hypotheses regarding student achievement. A statistical discriminant function analysis of data framed discussion on student outcomes, such that, scores on constructed/extended response items at grades 3 and 4 were significant predictors of overall proficiency at grade 5, (*p*< .05). Additionally, subcategories identified at grades 3 and 4 were significant predictors of student proficiency in grade 5. Ultimately, if performance on CR items is statistically related to students’ future mathematic achievement on all formats, educators should consider how constructed/extended response formats inform pedagogy, curriculum, and classroom methodology for mathematics at the elementary school level (Schoenfeld, 2002). Implications for instructional and assessment alignment are examined. The discussion centers on the predictive nature of constructed/extended response questions and the specific nature of subcategories of constructed/extended response at each grade level.

Theoretical background of the study

“Assessment in school is not a casual affair; not for the school, district, or state that must demonstrate adequate yearly progress (AYP) as part of public accountability and not for students working to meet high performance standards.” (Luke & Schwartz, 2010, p. 1) According to Darling-Hammond et. al. (2010), assessments are embedded in the curriculum; they influence the day‐to‐day work of teaching and learning and focus the use of knowledge to methods of solving problems.

Educators are experiencing a relentless pressure to show their effectiveness through student performance on standardized achievement tests. According to Valli, et. al. (2008) high-stakes tests lead teachers to stray from the qualities of good teaching including setting aside learner-sensitive responses in favor of moving lessons forward and covering the identified content, reducing the cognitive challenge of the lessons they designed, and posing lower level questions. In a study by Au (2007) results indicated that high-stakes testing exerts significant amounts of control over the content, knowledge forms, and pedagogies at the classroom level. His study outlined high-stakes tests that “encourage curricular alignment to the tests themselves”, narrowing the curricular content to tested subjects, to the detriment or exclusion of non-tested subjects. The findings of his study further suggested that the structure of the knowledge itself also changed to meet the test-based norms as “content was increasingly taught in isolated pieces and often learned only within the context of the tests themselves.” He stated that “Both content contraction and the fragmentation of knowledge, pedagogy” was also implicated, as “teachers increasingly turn to teacher-centered instruction to cover the breadth of test-required information and procedures.” (p. 263). Pedulla (2003) and Clarke (2002) surveys of educators confirm that the [NCLB] model promotes teaching to the test and a narrow curriculum (Guisbond & Neill, 2004).

Although Falkner et. al. (2006), in his study of testing in Kentucky classrooms, stated respondents in their study found the state assessment to be a beneficial instructional tool that provided focus for instruction and kept teachers “on track,” most teachers felt state assessment mandates had negatively affected their use of instructional time and selection of instructional strategies.

Not all researchers and educators agree with these perspectives, for example as Yeh (2005) found that teachers in Minnesota reported that their pedagogy was not negatively affected by high-stakes tests because they said the tests there were well designed and did not promote drill and rote memorization. Another example comes from Hillocks (2002) where investigations led to the belief that different latent traits were measured with various formats. As Darling Hammond (2010) suggests:

“U.S. tests rely primarily on multiple‐choice items that evaluate recall and recognition of discrete facts, examinations in most high‐ achieving countries use primarily open‐ended items that require students to analyze, apply knowledge, and write extensively. Furthermore, these nations’ growing emphasis on project‐based, inquiry‐oriented learning has led to an increasing prominence for school‐based tasks, which include research projects, science investigations, development of products, and reports or presentations about these efforts.”

When looking through the lenses of Hillcock, Yeh and Darling–Hammond, the key element in their discussion is the construction of the test itself. The real issues is not the item format per se, but the use and impact of item format on the overall design of classroom instruction including the analysis of students’ understanding which then impact teachers’ next steps. According to Kang, et.al. (2007), test format and corrective feedback modify the effect of testing on long-term retention. “The more demanding the retrieval processes in testing engendered by an intervening short answer test, the greater the benefit to final retention.” The practical application of his findings was that regular short answer quizzes with feedback might have been effective in enhancing student learning than repeated presentation of target facts or taking a multiple choice quiz (p. 528).

Those that content MC questio9n might predict outcomes in the same manner, ignore many researchers perspectives that multiple-choice tests do not provide information on the thinking processes of the respondents, which are important to differentiating between deficiencies in thinking and differences in background beliefs and assumptions (Arter & Salmon, 1987; Ennis, 1993; Norris, 1990). According to Tao;

“MC questions give a misleading impression of what mathematical problem solving is, and how one should go about it.  In actual mathematical research, problems do not usually come with a list of five alternatives, one of which is correct; often, figuring out what the potential, plausible, or likely answers could be, or even what type of answers one should expect or whether one should ask the question at all, is as important as actually identifying the correct answer.  Multiple choice quizzes also tend to reward quick-and-dirty or sloppy approaches to problem solving, as opposed to careful, deliberate, and nuanced approaches; in particular, such quizzes tend to encourage the mindless application of formal rules in order to arrive at an answer, without devoting much thought as to whether these rules are actually applicable for the problem at hand. (Tao, 2008).

Test developers have attempted to alleviate this by asking respondents to provide reasons for their answers, incorporating open-ended free response questions instead (Ennis, 1993; Halpern, 1994). A review of the California State Department of Education’s report on open-ended questions, *A Question of Thinking*, shows that most students lack opportunities to express mathematical ideas in writing, with fewer than 25% able to write completely about any of the problems given (Stenmark, 1989). Part of effective instruction is giving students opportunities to explain their thinking in writing, using proofs, multiple steps, organizers and written sentences. we now know that in order to learn mathematics, students must learn to communicate mathematically (NCTM 2000). This means listening, speaking, reading, and interpreting. It means explaining how a problem is solved, and explaining the problem and its solution using a variety of representations: words, symbols, graphs, charts, visuals, models, and manipulatives (Leiva, 1995).

Additionally, Glover (1989) concluded also that the more complete the retrieval operations during the intervening test, the greater the benefit to final memory performance. Kang also suggests (2007), “If the processes engaged during memory retrieval are crucially responsible for the testing effect, then one might expect that the more demanding or effortful the retrieval during a test, the better that material will be remembered later.” The “retrieval demands hypothesis” suggests that an intervening short answer (SA) test resulted in better performance on the final test than an intervening multiple choice (MC) test, regardless of whether the final test was MC or SA format. These findings showed that of the two test formats often used by teachers, the SA test was more beneficial for long-term retention than restudying. Additionally, results indicated test that item format and corrective feedback modulated the testing effect. Completing a SA test was found to boost final test performance more than additional focused exposure to test-relevant information, if corrective feedback was provided to improve poor initial test performance. Although taking an MC test improved final retention relative to completing no test after study, performance was not significantly different from receiving additional exposure to test-relevant information” (p. 544).

Authors have argued that if a conceptual base of knowledge in mathematics is limited, children may fail to develop skills which may prove necessary for future success (O’Neil & Brown, 1998). Skills, such as critical thinking or problem solving, which are paramount to achieving success may be undermined long before students reach secondary grade levels (Carnine, Jones, & Dixon, 1994), suggesting that narrowly aimed assessment reform is not sufficient (Hirsch, Koppich, & Knapp, 2001). Because choices students make in mathematics are based on the most commonly used instruction techniques; a unitary method of representation is often used exclusively when solving particular types of problems (Tsamir & Almog, 2001). As a result, assessment reformers, focusing on evaluation of student knowledge, stress the need for integration between assessment and instruction (Shepard, 2000). Researchers (Clark, 2002; Pedulla, 2003) who surveyed educators confirm that the No Child Left Behind model promoted teaching to the test and development of narrowed curriculum (Guisbond & Neill, 2004). This can be seen in in traditional testing, where selected response, such as multiple choice (MC) exams, are common, due to the relative ease and simplicity of implementation. However, selecting the correct response can be a false indicator of knowledge (Dufresne, 2002), as MC exams may be more reflective of shallow factual memorization, or one's ability to take an exam (e.g. test smarts) as opposed to actual knowledge (Stiggins, 2005). According to Tomlinson (2007), this format provides little information about problem solving abilities and other critical information teachers use to make informed judgments regarding appropriate instructional strategies and practices.

One way in which assessment and instruction may be better aligned to develop more in-depth understandings and knowledge is through the use of appropriate testing items, i.e. open-ended/constructed and short answer response items. Although much remains to be learned about test item formatting and its impact on classroom instruction, using these formats has been linked to promoting models of teaching and learning which facilitate in-depth understanding (Lane, 2004). Alternatively, the use of open-ended/constructed and short answer response items may allow educators to gain unique diagnostic insight into students’ thinking, while providing opportunities for students to exercise personal agency to enhance learning (Bahlmann & Walter, 2006; Birenbaum & Tatsuoka, 1987).

*Methodology*

This evaluation utilized data from one school in one upstate New York State school district. The data encompassed longitudinal archival test results from approximately 300 students in grades 3 through 5, from a school identified in the New York State Report Card (2009) as an elementary level school in an urban or suburban school district with high student needs in relation to district resources. Comparable schools in this group were in the middle range of student needs for elementary level schools in these districts.

A correlational multivariate research design using *a priori* data was utilized through a discriminant analysis in order: 1) to investigate which content-specific variables from grades three and four best predict fifth grade proficiency level; 2) to determine the degree of association of third and fourth grade content variables with 5th grade outcomes and 3) to outline the generation of weights that allow for the establishment of predictive functions. Additionally, the analysis used components of two assessment tools to assess the dependent and independent variables based on mathematics scaled scores obtained from the administration of either the Test of New York State Standards (TONYSS; created by Riverside Publishing Company, 2006) and the New York State Testing Program, Mathematics Tests from 2003 -2009.

*Results and Conclusions*

Overall, the data analysis in this study suggested that open-ended, extended/ constructed response questions in grades three and four can predict proficiency grouping on future fifth grade mathematics test. All content constructs (grade 3 and 4 statistics and probability, grade 3 and 4 number sense, numeration and operations, grade 3 and 4 geometry, grade 3 and 4 algebra and grade 4 measurement) appeared to be important when attempting to discriminate between the proficiency levels in fifth grade student outcomes. More specifically, in third grade, statistics, probability and geometry were the strongest predictors of fifth grade proficiency level. These constructs require students to understand mathematics through the collection, organization, displaying, and analyzing of data. Students make predictions that are based upon data analysis and understand and apply concepts of probability (NYSED, 2010, p. 26). In the case of geometry, mathematics is realized through visualization, spatial reasoning and analyses of characteristics and properties of geometric shapes (citation).

Additionally students formally and informally identify and justify geometric relationships, apply transformations, use symmetry, and coordinate geometry to analyze problem solving situations (NYSED, 2010, p. 26). In grade 4 number and numeration and algebra are the most predictive sub-categories. Students understand numbers through multiple ways of representing numbers, relationships among numbers, and number systems. Students glean meanings of operations and procedures and how they relate to one another and compute accurately to make reasonable estimates when dealing with number sense and numeration (NYSED, 2010, p. 26). In algebra, students represent and analyze algebraically problem solving situations; perform algebraic procedures accurately; recognize, use, and represent algebraically patterns, relations, and functions (NYSED, 2010, p. 26).

These constructs are supported in classrooms through activities focused on specific bodies of knowledge and mental processes associated with each sub-category area and the grade level concepts. While the content may look similar across grade levels the mental processes used at each grade level change and become more sophisticated as the grade level increases. In comparing the third and fourth grade predictive value of the subcategories, the differences may be due to the specific content taught and the level or depth of the concept within each category.

Overall, however, outcomes supporting the predictive nature of open-ended questions add to researchers’ contention that the use of open-ended constructional formats improves the development of conceptual understanding and builds long-term conceptual learning (Gray & Tall 2002). Additionally, it hints at a need to address the issues of testing preparation, the impact of testing on classroom instruction, the discussion of what constitutes in-depth understanding in mathematics learning, how it should be supported, and how it is practiced (Moon & Schulman, 1995; Kuechler & Simkin, 2010). Additionally, these outcomes support research that suggests the whole curriculum can be framed with an awareness of the abstract, evoking complex thinking skills (Gray & Tall 2002; Martinez, 2010).

*Implications*

The idea of open-ended items promoting higher order thinking skills is not a new one and there has been world-wide demand for curriculums to engage students in learning activities that nurture critical thinking skills. To understand whether we have been successful in answering such demand, the development of suitable assessment of critical thinking is crucial. According Ku’s (2009) in-depth study of item formats and their cognitive and dispositional value for critical thinking, open-ended response formats would capture more of the cognitive ability and disposition aspects of critical thinking than multiple-choice response formats would (Ku, 2009, p. 71). Additionally, researchers contend that testing impacts how teachers teach by altering instructional practices and the philosophy behind methods they use. (Barksdale-Ladd & Thomas 2000; Jones & Johnston, 2002; Wideen et al 1997; Yarborough, 1999). According to Kazemi, (2002), assessment can be a source of insight into student learning, but it requires that we pay close attention to children’s thinking. The research behind this study of open-ended mathematics test items suggests that item format is not irrelevant to math achievement measured in elementary grades and can inform effective instruction of mathematics.

Although this research suggests that the item formats can help inform instructional methods, the specific content strands and the reason behind one predicting outcomes more than another, still need to be investigated in more depth. While these analyses may be far from inferring support for the development of more complex conceptual thinking and mathematical awareness in classrooms, the first step in this direction is the investigation of the connection and explanation of the relationship between open-ended question, classroom instruction and student achievement. Additionally, it suggests that we need to investigate in more depth, the skills learned when using multiple choice items on tests at this early developmental age as a per ponderous use of multiple choice type formats will continue to exacerbate limitations on children’s understanding of mathematics.

References

Arter, J.A. & Salmon, J.R. (1987). *Assessing higher order thinking skills: A consumer’s guide.*

*Portland*: Northwest Regional Educational Laboratory.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis.

*Educational Researcher, 36*(5), 258-267.

Bahlmann & Walter, (2006) Exploring the open-response task as a tool for assessing the understanding of fifth grade students in the content area of fractions. In Alatorre, S., Cortina,

Barksdale-Ladd, M., & Thomas, K. (2000). What’s at stake in high stakes testing?

Journal of Teacher Education, 51, 384-397.

Birenbaum, M., & Tatsuoka, K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11(4),* 385-395

Carnine, D., Jones, E., & Dixon, R. (1994). Mathematics: Educational tools for

diverse learners. *School Psychology Review, 23*(3), 406-427.

Clarke, D.J. (2002). Constructed Pedagogies. In L. Bazzini & C. Whybrow Inchley (Eds.), Mathematical Literacy in the Digital Era: Proceedings of CIEAEM 53. Milano: Ghiseti E Corvi, 243-252.

Darling-Hammond, Pecheone, Jacquith, Walker, Wei (2010). *Developing an internationally*

*comparable balanced assessment system that supports high-quality learning*. National Conference on Next Generation Assessment Systems Content Provided by the Center for K – 12 Assessment and Performance Management. Educational Testing Service,

Dufresne, W. J., Gerace, L. Gerace, W.J. Making sense of students’ answers to multiple choice questions. *The Physics Teacher*. *40*, 174 – 180.

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179-186.

Finbarr S. C., & Kelly A. E. (2003). Issues in High-Stakes Testing Programs. *Theory Into Practice, 42,* 12-18.

Falkner, S. & Cook, C. (2006). Testing vs. Teaching: The Perceived Impact of Assessment Demands on Middle Grades Instructional Practices, *Research in Middle Level Education, RMLE Online*. *29*(7),ISSN 1084-8959

Glover, J. A. (1989). The ‘‘testing’’ phenomenon: Not gone but nearly forgotten. *Journal of*

*Educational Psychology*, 81, 392- 399. In Kang, S.H., McDermott, K. B., Roediger, H. L. III, (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19* (4/5), 528-558.

Gray, E., & Tall, D. (2002). *Abstraction as a natural process of mental compression*. In

Cockburn, A., & Nardi, E. (Eds), Proceedings of the 26th Conference of the

International Group for the Psychology of Mathematics Education, Norwich, UK,

*1*, 115–120.

Guisbond, L., & Neill, M. ( 2004). Failing our children: No Child Left Behind undermines quality and equity in education. Clearing House, 78(1), 12-16.

Halpern, D. F. (Ed.). (1994). *Changing College Classrooms: New Teaching and Learning*

*Strategies for an Increasingly Complex World*. San Francisco: Jossey-Bass.

Hillocks, G.,Jr. (2002). The Testing trap: How states writing assessments control learning. New

York: Teachers College Press.

Hirsch, E., Koppich, J., & Knapp, M. (2001). *Revisiting what states are doing*

*to improve the quality of teaching: An update on patterns and trends*.

Seattle, WA: Center for the Study of Teaching and Policy, University of

Washington.

Hollingworth, L., Beard, J. J., & Proctor, T. P., (2007.) An investigation of item type in a standards-based assessment. *Practical Assessment, Research and Evaluation*. *12*(18).

Jones, B. D., & Johnston, A.F. (2002, April). The effects of high-stakes testing on

instructional practices. Paper presented at the 2002 annual meeting of the

American Educational Research Association, New Orleans

Kang, S.H, McDermott, K., Riediger, H.LK. III (2007) Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19* (4/5), 528-558

Kazemi, E. (2002). Exploring test performance in mathematics: The questions children’s

answers raise*. Journal of Mathematical Behavior*, *21*, 203–224.

Ku, K. (2009). Assessing students’ critical thinking performance: Urging for

measurements using multi-response format*. Thinking Skills and Creativity, 4(1),*  70–76

Kuechler, W., & Simkin, M. (2010). Why is performance on multiple-choice tests and

constructed-response tests not more closely related? Theory and an empirical test.

*Decision Sciences Journal of Innovative Education, 8*(1), 55-73.

doi:10.1111/j.1540-4609.2009.00243.x

Luke, S. & Schwartz, A. (2007). Assessment and Accommodation. *Evidence for Education*. 2(1), Resources updated, October 2010. National Dissemination Center for Children with Disabilities (NICHCY), Washington, DC.

Martinez, M. E. (2010). *Learning and cognition: The design of the mind*. Upper Saddle River, NJ: Pearson Education, Inc.

Moon, J., & Schulman, L. (1995). *Finding the connections: Linking assessment,*

*instruction, and curriculum in elementary mathematics*. Portsmouth, NH:

National Council of Teachers of Mathematics (2000). Principles and Standards for School Mathematics. Reston, VA: Author.

NAEYC & NCTM (National Council of Teachers of Mathematics). 2002. Math experiences

that count! Young Children 57 (4): 60–62.

New York State Education Department (2009). *The New York State Report Card*. Retrieved from <https://reportcards.nysed.gov/>

New York State Education Department (2010). *Learning Standards of New York State.* Retrieved from <http://www.p12.nysed.gov/ciai/standards.html>

O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math:

Assessment on metacognition and affect*. Applied Measurement in Education,*

*11*(4), 331-351. doi: 10.1207/s15324818ame1104\_3

Pedulla, J.J. (2003) State-mandated testing: What do teachers think? *Educational Leadership* *61(3),* 42-46.

Riverside Publishing Company (2006). *Test of New York State Standards* (TONYSS).

Educational Testing Service, 1998 in Martinez, 1999)

Schofield, J W. (2002). Increasing the generalizability of qualitative research. In Huberman, A. M. & Miles, M. B. (Eds.). The qualitative researcher’s companion (171-203). Thousand Oaks,CA: Sage

Shepard, L. (Winter, 2000). Why is teaching to the test a bad thing? State educational

leaders. 7-8

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20,* 8-11.

Smith, M. L., Edelsky, C. E., Draper, K., Rottenberg, C., & Cherland, M. (1989). *The role of external testing in elementary schools*. Los Angeles: University of California, Center for the Study of Evaluation.

Stecher, B. M., & Borko, H. (2002). Integrating findings from surveys and case studies:

Examples from a study of standards-bases educational reform. Journal of Education Policy,

17(5), 547–569.

Stenmark, J. K. (1989) Assessment alternatives in mathematics, An overview of assessment techniques that promote learning. Berkeley: EQUALS and the California Mathematics Council.

Stiggins, R. J. (2005). Student-involved assessment for learning (4th ed.). Columbus, OH: Merrill, Prentice Hall.

Simon, E., Foley, E., Passantino, C. (1998). *Making sense of standards: Implementation issues and the impact on teaching practice.* Consortium for Policy in Education and Research for Action.

Riverside Publishing (2006). Standardized Test of New York State Standards (TONYSS). Riverside Publishing Company, Rolling Meadows, IL. Subsidiary of Houghton Mifflin Harcourt Publishing Company,

Tao, T. Mathematics, non-technical, opinion, teaching: On multiple choice questions in mathematics. Retrieved on January 25, 2013, from http://terrytao.wordpress.com/2008/12/14/on-multiple-choice-questions-in-mathematics/

Third International Mathematics and Science Study, U.S. mathematics, Boston, MA.

IEA (1999). Third International Mathematics and Science Study, U.S. Mathematics International Study Center Chestnut Hill, MA: TIMSS International Study Center,

Tomlinson, C. A. (2007). Learning to love assessment, *Educational Leadership,* 65

(4) 8-13.

Tsamir, P., & Almog, N. (2001). Students' strategies and difficulties: The case of algebraic inequalities. *International Journal of Mathematical Education in* *Science and Technology, 32,* 513-524.

Taylor, G., Shepard, L., Kinner, F., Rosenthal, J. (2002). A survey of teachers' perspectives on high-stakes testing in colorado: What gets taught, what gets lost. Los Angeles, CA: California University, Center for the Study of EvaluationVergnaud (1988)

United States Education Department (2010). A blueprint for reform: The reauthorization

of the elementary and secondary education act. Retrieved from

http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

Valli, L., Croninger, R. G., Chambliss, M., Graeber, A. O., & Buese, D. (2008).High-stakes accountability in elementary schools. New York, NY: Teachers College Press.

Wideen, M.F., O’Shea, T., Pye, I., & Ivany, G. (1997). High-stakes testing and the teaching of science. Canadian Journal of Education, 22, 428-444.

Yeh, S.S. (2006). Can Rapid Assessment Moderate the Consequences of High-Stakes Testing? Education and Urban Society, 39, 91-112.

Yarbrough, T.L. (1999). Teacher perceptions of the North Carolina ABC program and the relationship to classroom practices. Unpublished doctoral dissertation,

University of North Carolina at Chapel Hill.