Assessing Mathematical Proficiency
MSRI Publications
Volume **53**, 2007

Chapter 13
Task Context and Assessment

ANN SHANNON

Introduction

In this chapter, I will explore the impact of task context on assessment in mathematics. It is nontrivial to determine the understandings measured by a given assessment, so a close examination of some tasks and what they reveal is the main focus of this paper. Before considering these contemporary explorations, and in order to establish for the reader that the context of a mathematics task is indeed a salient feature, I will review findings from research in mathematics education and psychology. I will show that the role of context in mathematics assessment is a complex issue that involves much more than capturing the interest and harnessing the motivation of the student.

Background

Keeping it real. Over the past several decades, many different researchers and educators have pointed out the benefits of setting mathematical tasks in rich, attractive, and realistic contexts (e.g., [de Lange 1987; Freudenthal 1983]). Realistic contexts are generally regarded as referring to aspects of the "real" social or physical world as well to fictional, imaginary, or fairy-tale worlds. Specifically, there are no restrictions on the contexts that can be called realistic as long as they are meaningful, familiar, appealing, and morally appropriate for students. In the literal sense, it is not the degree of realism that is crucial for considering a context as realistic, but rather the extent to which it succeeds in getting students involved in the problem and engages them in meaningful thinking and interaction.

Realistic contexts are recommended for two main reasons. On the one hand, it is thought that a realistic context will facilitate student success by intrinsically motivating students and thus increasing the likelihood that they will make a serious effort to complete the problem. On the other hand, a realistic context may facilitate performance by helping students to make a correct representation of the problem and to formulate and implement a feasible solution strategy by activating the use of prior knowledge specific to that context that is helpful for understanding and solving the problem.

There is growing evidence, however, that realistic contexts can have a negative impact on the performance of students on mathematics tasks. For example, [Boaler 1994] analyzed the performance of 50 students from a school with a traditional approach to mathematics education on two sets of three questions intended to assess the same content, but set in different contexts. Interestingly, the girls from the school that Boaler studied tended to attain lower grades on an item that had been cast in the context of "fashion" than they did on isomorphic items that were cast either in an abstract context or even in a realistic context that was thought to be inherently less appealing to the girls (such as football — "soccer" in U.S. terms). According to Boaler, the girls' relatively poor performance on the fashion item was caused by the problem context distracting them from its deeper mathematical structure.

More recently, De Bock, Verschaffel, Janssens, Dooren, and Claes [2003] also found that context could have a negative impact on students' performance. These authors analyzed student responses on tasks set in a context that included Lilliputians (of Swift's *Gulliver*). The authors suggest that, just like the girls in Boaler's study, their students' emotional involvement with the Lilliputians may have had a negative rather than positive influence on student performance. Thus, at the same time as we observe textbook writers and test developers infuse their new curriculum and assessment materials with sought-after realistic contexts, current educational research evidence is making it increasingly clear that the underlying case for realistic contexts has neither been well made nor well understood.

Assessing reasoning. In the early 1970s, a flurry of investigations into the role of context in reasoning was motivated by Piaget's theory of formal operations. Briefly, Piaget's earliest rendering of the stage of formal operations described this level of thought as unshackled by either content or context of a problem. Instead, it was believed that with the onset of the *formal operational* stage of thinking, problem solvers were guided by propositional logic and the problem's structure, rather than its content or context.

The individual and combined work of Wason, Johnson-Laird, and others within the British and U.S. cognitive psychology community developed what is affectionately known today as the Four-Card Problem [Wason 1969]. This work showed very clearly that when people solve a problem they usually rely upon its contextual features rather than solving the problem by abstracting form from content as Piaget's initial theory of formal operations had suggested.

The problem was often presented as follows:

Four-Card Problem

Here are four cards. You know that each has a number on one side and a letter on the other. The uppermost face of each card is like this:

A P 6 3

The cards are supposed to be printed according to the following rule:
If a card has a vowel on one side, it has an even number on the other side.

Which among the cards do you *have* to turn over to be sure that all four cards satisfy the rule?

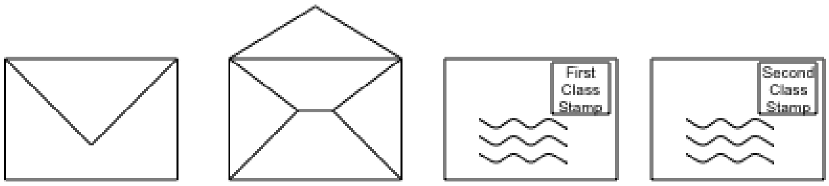
Before you read any further, try to answer the question. You probably decided that you needed to turn over the card with the A, because if the number on the other side turned out not to be even, you would have disproved the rule. In the 1970s, most people almost always got this right. Similarly, most people knew that they did not need to turn over the card with the P, because the rule says nothing about consonants. In the 1970s, however, the research participants were divided as to whether or not it was necessary to turn over the card with the 6. The correct answer is that you do not need to turn over the card with the even number. This is because it might have either a vowel or a consonant on the other side and neither card will violate the rule. With regard to the card with the number 3 on it, most of the research participants decided that they did not need to turn it over in order to check the rule. The correct answer, however, is that the card with the number 3 must be turned over to verify that the other side does not contain a vowel. If the other side contains a vowel then the rule will not be satisfied.

Due to the low incidence of correct responses on the Four-Card Problem, researchers created different versions of the task that employed thematic contexts. For example, during the 1970s in England and parts of Ireland, there were two main rates for mailing envelopes — first and second class, first class being more expensive. At that time, if you sealed your envelope you had to put a first-class stamp on it, but if your envelope was left unsealed then a second-class stamp would do. For some people living in England and parts of Ireland during that time, the context was real and relevant and thus an envelope version of the Four-Card Problem was created in order to assess reasoning in a real-world context [Johnson-Laird et al. 1972].

Four-Envelope Problem

Here are four envelopes. Each has a stamp on one side and each is either sealed or left unsealed.

The uppermost face of each envelope is like this:



The envelopes are mailed according to the following Post Office rule:
If an envelope is sealed then it must have a first-class stamp on it.

Which among the envelopes do you have to turn over to be sure that all four envelopes satisfy the Post Office's rule?

The Four-Envelope Problem is identical to the Four-Card Problem in structure. In the early research in Britain, participants found the envelope version of the task easy to solve and made few errors. Compared to the Four-Card Problem, the Four-Envelope Problem was almost a trivial exercise, because the solvers' understanding of the familiar context carried them to a successful conclusion. Thus, because of the context, participants were able to say that the sealed envelope and the one with the second-class stamp were the envelopes that needed to be checked. Participants also responded correctly that neither the unsealed envelope nor the one with the first-class stamp needed to be checked because if people wanted to waste a first-class stamp on an unsealed envelope then that was their business. The Post Office rule did not stipulate the kind of stamp that needed to go on an unsealed envelope nor did it say anything about sealing an envelope with a first-class stamp. Johnson-Laird et al. [1972] reported what they called a "thematic-materials effect," given the way in which the postal-rule context facilitated improved performance compared to the vowels and numbers context.

In the 1980s, members of both the U.S. and British cognitive psychology community had difficulty replicating the results of that particular experiment. Researchers in the U.S. could not replicate it because the postal regulations of the Four-Envelope Problem have never existed in that country [Griggs and Cox 1982]. British researchers could not replicate this effect when their participants were too young to have experienced the obsolete postal regulations (Golding, 1981, as cited in [Griggs and Cox 1982]). In response to the U.S. undergraduates' poor performance on the Four-Envelope Problem, Griggs and

Cox ingeniously invented their own thematic version. The context for this task is one that is undoubtedly as real and as relevant today as it was in the 1980s.

Four Drinkers Problem

You are in charge of a party that is attended by people ranging in age. The party is being held in a state where the following law is enforced:

If you are under 21 you cannot drink alcohol.

Your job is to make sure that this law is not violated.

Understandably, you want to check only those people who absolutely need to be checked. At one table there are four people drinking. You can see the IDs of two of these people: One is under 21 and one is older than 21. You do not know what these two are drinking. You do however know what the other two people at the table are drinking: One is drinking soda and the other is drinking beer. You cannot see the ID of either of these two people. So to summarize your problem:

Under 21	Over 21	Drinking soda	Drinking beer

Of these four, who do you need to check in order to make sure that the law is not broken?

Within this context, you can probably see at a glance that you need to check both the drink of the person who is under 21 and the ID of the person who is drinking beer. But you probably would not think of checking the drink of a person who is older than 21, because the law does not say anything about the drinking habits of a person who is over 21. Similarly, you would not think of checking the ID of a person drinking soda because the law does not have a legal age for drinking soda.

I have chosen to take the reader through just a few of the highlights of this now decades-old thread of psychological research because the original Four-Card Problem and its subsequent recastings cogently "problematize" the issue of task context and assessment. These highlights clearly show that context can aid or impede the solver, and also show that context can sometimes change a task so substantially so as to lead one to ask, "Is this still a math task?"

From the highlights, you can see that we have a well-tested example of three tasks with the same mathematical structure, which appear to be solved quite differently. For most people, the Four-Card Problem is essentially a problem in pure logic. On the other hand, for some people, the envelope and drinking age versions of the problem may not be problems in pure logic. In the latter variations, the success of the problem solver has been shown to depend heavily on whether the context of the problem is familiar and meaningful to the solver. Given sufficient familiarity with context, it seems that the solver can be carried along by the meaning of the context, and so is prevented from making the logical errors that trip up participants on the Four-Card Problem.

In the envelopes and drinking age scenarios, the familiarity of the situation rather than the mathematical structure facilitated success. But with regard to assessment, what does participant performance on these tasks tell us about the participants' understanding of mathematical logic? More specifically, what do these performances tell us about what the participants might have learned about propositional logic or about analysis of propositions of the form "If P then Q"? Interestingly, the cognitive researchers found that prior success and experience with familiar contexts such as envelopes and drinking age did not readily transfer to the abstract Four-Card Problem involving vowels and consonants [Cox and Griggs 1982; Johnson-Laird et al. 1972; Griggs and Cox 1982; 1993]. Thus, it seems that if our aim were to assess learning about propositional logic, it would not make sense to deploy a thematic context. When it comes to propositional logic, thematic context can lead to false-positive or false-negative results; see, for instance, [Cox and Griggs 1982].

This discussion is not meant to suggest that task context has nothing to with the assessment of mathematics. The issue of assessment is just far too complex for that [Boaler 1994]. To the contrary, I will argue that assessment of important mathematics can be facilitated by tasks with appropriate real-world context, provided that the task context and the mathematics to be assessed are sufficiently integrated. This can be accomplished if problem solvers are invited to use the context to demonstrate some aspect or aspects of their mathematical prowess.

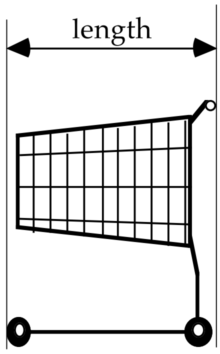
Thus, the relevant question that this cognitive research raises in relation to the more contemporary problem of teaching and learning school mathematics is: How can familiar, real, and relevant contexts be used effectively to assess mathematics? Some insight into this question is afforded by another series of studies that I carried out for the Balanced Assessment and New Standards projects [Shannon and Zawojewski 1995; Shannon 1999; 2003].

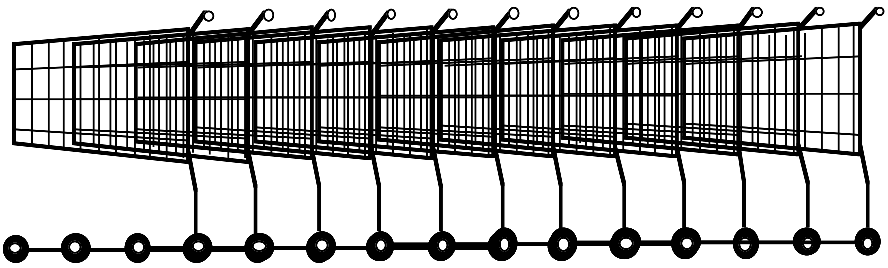
Beyond Interest and Motivation

As part of my work for the Balanced Assessment Project (a task development project funded by the National Science Foundation) and later for New Standards, I conducted a series of mini-studies focused on a group of three very similar tasks involving linear functions in real-world contexts [Shannon 1999]. The three tasks were called Shopping Carts, Shopping Baskets, and Paper Cups. In each task, students were presented with diagrams of common objects that can be nested when stacked, and were asked to measure the diagrams and develop linear functions describing how the length or height of a stack would vary with the number of objects in the stack.

Shopping Carts

The diagram shows a drawing of a single shopping cart. It also shows a drawing of 12 shopping carts that have been nested together. The drawings are 1/24th real size.



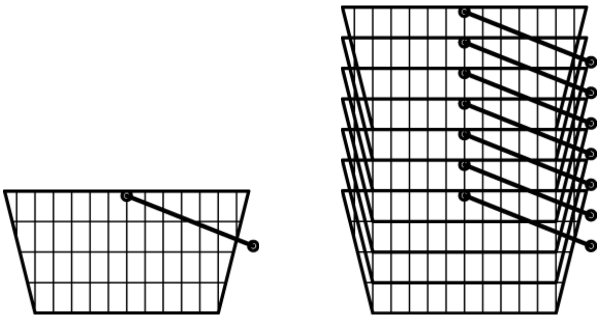


Create a formula that gives the length of a row of nested shopping carts in terms of the number of carts in that row.

Define your variables and show *how* you created your formula.

Shopping Baskets

The diagram [next page] shows a drawing of a single shopping basket. It also shows a drawing of 7 shopping baskets that have been nested together. The drawings are 1/10th real size.



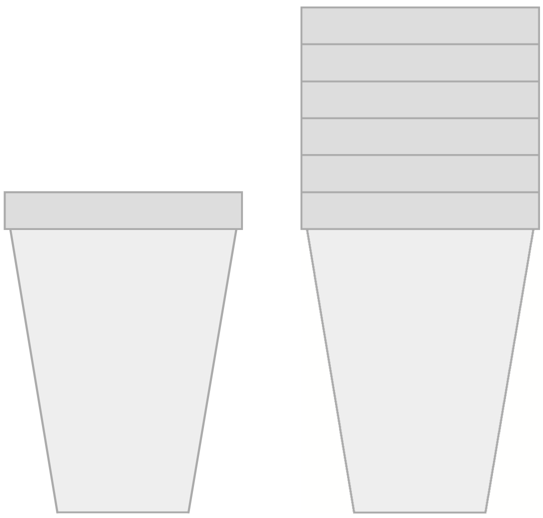
1 basket 7 nested baskets

Create a formula that gives the height of a stack of shopping baskets in terms of the number of baskets in a stack.

Define your variables and show *how* you created your formula.

Paper Cups

The diagram shows drawings of one paper cup and of six paper cups that have been stacked together. The cups are shown half size.



1 cup 6 stacked cups

Create a formula that gives the height of a stack of cups in terms of the number of cups in the stack.

Define your variables and show *how* you created your formula.

The three tasks described here are clearly all variants of a common theme, and they tap into the same area of mathematics — linear functions and arithmetic sequences (linear functions of positive integers). Despite their common mathematical structure, the three tasks proved to present three separate levels of challenge for most students. Shopping Carts was the most difficult, followed by Shopping Baskets, and then Paper Cups. To appreciate the nature of the challenges that these tasks offered to students, I analyzed the responses from comparable groups of students on Shopping Carts and Shopping Baskets, then analyzed responses from comparable groups of students on Shopping Baskets and Paper Cups.

Shopping carts versus shopping baskets. In using the pictures of the stacks to make the necessary measurements, I have found that students are considerably more successful in using the diagram of the baskets than they are in using the more detailed diagram of the carts. For example, it has seemed that the wheels and handles of the carts present students with a much more complicated diagram to work with, when compared with the relatively more straightforward picture of the baskets. Thus, the authenticity of the diagram of the stack of carts presented students with many extraneous details, and students had to identify those aspects of the structure of the stack of carts that were relevant to the problem and those that were not relevant.

There are also scale differences between these two versions of the task. The carts were drawn as 1/24 of the actual size, while the baskets were drawn to 1/10 scale. Students have proved to make fewer computational errors using the scale factor of 1/10 for Shopping Baskets than when using the scale factor of 1/24 needed for Shopping Carts, and they have also seemed better positioned to know how to use this information in the case of the baskets than in the case of the carts.

Finally, there are differences of orientation. The length of the stack of carts increases in a horizontal direction while the height of the stack of baskets increases in a vertical direction, and it has seemed as a result that students are better able to visualize the increasing stack of baskets than they do the increasing stack of carts. While working on Shopping Baskets, for example, students were observed gesturing with their hands as if to show how the height of a stack of baskets might increase with increasing number, but no similar student actions were observed when students were working on Shopping Carts.

Shopping baskets versus paper cups. The baskets were drawn as $1/10$ of the actual size and the cups were drawn as $1/2$ of the actual size. There was no evidence that either of these scale factors provided a greater challenge to students. Similarly, students were no more nor less successful in using the picture of the stack of baskets to make measurements than in using the picture of the stack of cups. Both the stack of baskets and the stack of cups increased in a vertical direction. Thus, it seemed that the stack of baskets presented no more visual or measurement-related complications than did the stack of cups.

Despite these similarities, however, students were more successful in expressing the height of a stack of cups in terms of the number of cups in the stack than in expressing the height of a stack of baskets in terms of the number of baskets in the stack. This has appeared to be because the cups were depicted as having a discernible lip and base, and students have then found it fairly straightforward to decompose the height of the stack of cups into that of "one base and n lips." The structure of the baskets does not invite a similar decomposition, probably because the part of each basket that protrudes above the previous basket cannot readily be depicted as a separate entity, and cannot easily be named or conceptualized in the same way as the lip and base of the cup can be identified and named. Thus, students have to think of the stack of baskets as one basket and $n-1$ "stick outs." This small difference in the structure of the two stacks has seemed to make for a much larger impact on the complexity of the algebraic prowess that is needed to create a formula for each stack.

With the cups, students could finesse having to deal with $n-1$ lips and could instead write the formula as $h = nl + b$ (where h represents the height of the stack of cups, n represents the number of cup-lips, and l and b represent the actual measurements of the cup-lip and cup-base, respectively).

With the baskets, students cannot finesse having to deal with $n-1$ "stick outs" and instead have to find their way to expressing the formula as $h = p(n-1) + B$ (where h represents the height of the stack of baskets, n the number of baskets, B the actual height of a basket, and p the amount that each basket protrudes above the one below). There is no doubt that this difference in challenge is not trivial to students whose grasp of algebra is still fragile and not yet flexible.

When I present these findings at professional development workshops or at conferences, participants invariably attribute the relative success of students on Paper Cups to the familiarity of a paper cup in the everyday life of students. Workshop participants expressed the view that it made sense that a paper cup would be more familiar to students than a shopping basket. On the other hand, the widely held view that a familiar context will facilitate success on a mathematics task may lead many participants to interpret the finding that students are more successful on Paper Cups as evidence in and of itself that the paper cup must be more familiar to students than is the shopping basket.

My analysis of the student work suggests that the issue of the relative familiarity of a paper cup and a shopping basket is something of a red herring. It does however allow us to glimpse at the firmly held but perhaps erroneous beliefs that are held about the role of task context in the learning of mathematics. Based on my analysis of student work, I want to suggest that it is the specific geometry of the stack of cups that facilitates success with this version of the problem, rather than students' assumed greater prior experience with this everyday object. From the student work, it has seemed that it is simply easier in the case of the cups for students to translate from a visual representation (the diagram of the stack) to an algebraic representation (a formula). In later versions of these tasks, I have asked students to engage in a variety of related activities: express the height of the stack in terms of the number of cups or baskets in a stack; make a graph of that relationship; find the slope of the associated line; and finally, interpret the slope in terms of the initial situation. Students have proved much more successful in interpreting the slope when it represents the increase in height per cup (length of a cup-lip), as compared to the situation when it represents the increase in height per basket (length of the part of the basket that protrudes above the previous basket). Again, it seems reasonable to suggest that this is because (unlike "the part of a shopping basket that protrudes above each previous basket"), the cup-lip is a tangible object that can be named and understood with relative ease. There is clearly also a visualization aspect. The height of a cup lip is a vertical line segment on the diagram, but there is no line segment on the diagram to show the height of a basket's protrusion.

Discussion

The responses of students to tasks such as Shopping Carts, Shopping Baskets and Paper Cups suggests that the importance of these tasks for teaching and learning mathematics lies not in their authenticity or their familiarity for students, but in the opportunities that each of these structures provides to students in translating among different representations and in affording students the opportunity to engage in mathematical abstraction. Thus, I am suggesting that it is the geometry of the various stacks that make these important mathematics tasks, rather than the fact that the structures comprise mundane, and more or less familiar, or everyday objects.

It is interesting to note that tasks of this caliber proliferate in the curriculum of "Railside," the astoundingly successful urban high school mathematics department described by [Boaler 2004]. Boaler notes that the teachers in this highly successful mathematics department do not select the context of mathematics tasks so as to promote equity, or to relate to the students' own culture — rather, they select mathematics tasks so as to promote abstract mathematical discussions.

The capacity of a task to promote abstract mathematical discussion and thus precipitate learning depends on how the task is used. Boaler's research makes it clear that the effectiveness of the mathematics tasks assigned in Railside is due largely to the high expectations that teachers place on students as they work to complete the tasks. Specifically, Boaler's research, reports from the Railside teachers, and classroom visits to Railside all suggest that the Railside teachers' demand that all students engage in mathematical *justification* is instrumental in precipitating learning. For example, [Boaler 2004] discusses a video clip where students are asked to find the perimeter of a simple structure created using Lab Gear.[1] The structure is by no means intrinsically interesting, nor would it be familiar to the students in Railside's sheltered algebra class. However, working in groups and with varying degrees of difficulty, the students find that the perimeter is given by the expression $10x + 10$ or its equivalent. Boaler's video data show that the teacher is not satisfied by this correct answer and insists that each student explain in terms of the structure "where the 10 is." The teacher's tenacity is noteworthy, she persists until each student has justified the algebraic expression in terms of the structure created using Lab Gear materials. Thus, while the initial task simply asks students to represent the Lab Gear structure in terms of an algebraic expression, the teacher asks the student to go further, and to interpret the algebraic expression that they have created in terms of the underlying geometry. This approach provides a means for the teacher to ensure that the task provides the opportunity for students to translate among multiple representations, and challenges the students to do so in ways that are far from perfunctory. This short video clip communicates a clear sense that learning has taken place in the group, because students who struggle with the teacher's demands can be seen going on to tackle and accomplish subsequent, more difficult tasks with enthusiasm and confidence.

The examples discussed by Boaler make it clear that a task's capacity to precipitate student learning of mathematics will be highly dependent on how it is used in the classroom and on the particular efforts that are made to keep the cognitive demands of the task high [Schoenfeld 1988; Stein et al. 2001; Stein et al. 1996] and promote mathematical abstraction [Boaler 2004]. It is not an overstatement to say that if its cognitive demands are not maintained, the mathematics task — contextualized or not — will function no better than a series of "drill sheets." In particular, tasks with the potential to be worthwhile can be

[1] Lab Gear is a manipulative for algebra designed by Henri Picciotto.

rendered worthless if students are not given the opportunity to grapple with the intrinsic complexities of the underlying mathematics [Ball and Bass 2003].

This particular idea is one that is often lost in what seems to be a rush to wrap assessment tasks in a context in order to motivate students. Unfortunately, the seemingly firmly held belief that task context is a good thing often leads directly to poor quality assessment tasks. Consider this contemporary grade 6 example from a state testing program:

> There are 30 pencils left at a store after Shilo buys a certain number of pencils, p.
>
> Delia buys 4 times as many pencils as Shilo. The expression below shows the number of pencils remaining at the store after Delia buys her pencils.
>
> $$30 - 4 \times p.$$
>
> How many pencils remain at the store if Shilo bought 3 pencils?
>
> A. 14 B. 18 C. 78 D. 104

Among the difficulties with this task are the following:

(a) The mathematics itself is low-level: the problem simply asks, "What is the value of $(30 - 4 \times p)$ when $p = 3$?"

(b) The most important aspect of mathematizing, generating the formula, is not part of the task; in this connection see Chapter 7 in this volume.

(c) The linguistic complexity of the task far overshadows the mathematical complexity; see also Chapters 19 and 20 in this volume.

As point (a) indicates, this task can be completed by ignoring the context entirely and simply "plugging 3" into the given expression. This runs the risk of teaching students that context is not important — an unfortunate message to send to students since task context is extremely important when used correctly [Boaler 2004; Shannon 1999]. From an assessment standpoint tasks of this type, even if written as "free response" rather than multiple-choice questions, are problematic: if a student were to give the wrong answer it would be difficult to diagnose what caused the student to have problems. From an equity standpoint the task is problematic because the context clearly places extraneous reading demands on students. Thus it places unnecessary burdens on the shoulders of English learners and others who might find reading difficult. In sum, using context in a problem statement without examining its impact on students' problem-solving processes can be problematic.

Concluding Remarks

In looking for mathematics assessment tasks it is not necessary to look for a real-world context *per se*, but to look instead for the opportunities that the task provides for the student to formulate his or her own approach to the task, represent the solution in some appropriate and mathematically abstract form, and then interpret the salient components of the solution in terms of the initial task. The role of context is a complex and a subtle one, but there is no doubt that it plays a critical role in creating student access to worthwhile and important mathematics.

Acknowledgment

My sincere thanks to Meghan Shaughnessy for her help with reading drafts, locating references, and many other things.

References

[Ball and Bass 2003] D. L. Ball and H. Bass, "Making mathematics reasonable in school", pp. 27–44 in *A research companion to Principles and Standards for School Mathematics*, edited by J. Kilpatrick et al., Reston, VA, 2003.

[Boaler 1994] J. Boaler, "When do girls prefer football to fashion? An analysis of female underachievement in relation to realistic mathematic contexts", *British Educational Research Journal* **20**:5 (1994), 551–564.

[Boaler 2004] J. Boaler, "Promoting equity in mathematics classrooms: important teaching practices and their impact on student learning", paper presented at the International Congress of Mathematics Educators, July 2004.

[Cox and Griggs 1982] J. R. Cox and R. A. Griggs, "The effects of experience on performance in Wason's selection task", *Memory and Cognition* **10**:5 (1982), 496–502.

[De Bock et al. 2003] D. De Bock, L. Verschaffel, D. Janssens, W. Dooren, and K. Claes, "Do realistic contexts and graphical representations always have a beneficial impact on students' performance? Negative evidence from a study on modeling non-linear geometry problems", *Learning and Instruction* **13** (2003), 441–465.

[Freudenthal 1983] H. Freudenthal, *Didactical phenomenology of mathematical structures*, Dordrecht: Reidel, 1983.

[Griggs and Cox 1982] R. A. Griggs and J. R. Cox, "The elusive thematic-materials effect in Wason's selection task", *British Journal of Psychology* **73**:2 (1982), 407–420.

[Griggs and Cox 1993] R. A. Griggs and J. R. Cox, "Permission schemas and the selection task", *The Quarterly Journal of Experimental Psychology* **46A**:4 (1993), 637–651.

[Johnson-Laird et al. 1972] P. D. Johnson-Laird, P. Legrenzi, and M. S. Legrenzi, "Reasoning and a sense of reality", *British Journal of Psychology* **63**:3 (1972), 395–400.

[de Lange 1987] J. de Lange, *Mathematics, insight and meaning. Teaching, learning and testing of mathematics for the life and social sciences*, Utrecht: OW&OC, 1987.

[Schoenfeld 1988] A. H. Schoenfeld, "When good teaching leads to bad results: The disasters of 'well taught' mathematics classes", *Educational Psychologist* **23** (1988), 145–166.

[Shannon 1999] A. Shannon, *Keeping score*, National Research Council: Mathematical Sciences Education Board, Washington, DC: National Academy Press, 1999.

[Shannon 2003] A. Shannon, "Using classroom assessment tasks and student work as a vehicle for teacher professional development", in *Next steps in mathematics teacher professional development, grades 9–12: Proceedings of a workshop*, Washington, DC: National Academy Press, 2003.

[Shannon and Zawojewski 1995] A. Shannon and J. Zawojewski, "Mathematics performance assessment: A new game for students", *Mathematics Teacher* **88**:9 (1995), 752–757.

[Stein et al. 1996] M. K. Stein, B. Grover, and M. Henningsen, "Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms", *American Educational Research Journal* **33** (1996), 455–488.

[Stein et al. 2001] M. K. Stein, M. Smith, M. Henningsen, and E. Silver, *Implementing standards-based mathematics instruction: A case book for professional development*, New York: Teachers College Press, 2001.

[Wason 1969] P. C. Wason, "Regression in reasoning?", *British Journal of Psychology* **60**:4 (1969), 471–480.