Performance and Authentic Assessment, Realistic and Real Life Tasks: A Conceptual Analysis of the Literature

Torulf Palm

EM No 39, 2001

ISSN 1100-696X
ISRN UM-PED-EM--39—SE

Performance and Authentic Assessment, Realistic and Real Life Tasks: A Conceptual Analysis of the Literature[1]

Torulf Palm

Abstract

Performance assessment, authentic assessment, and realistic and real life tasks are recurrent terms in the literature on education. They have all been given a number of different meanings and unclear definitions and are in some publications not defined at all. The uncertainty of meaning causes difficulties in interpretation, scrutiny, and criticism of publications and may be damaging for the credibility of the use of these concepts in education reform. This paper reviews the meanings attached to these concepts in the literature and describes the similarities and wide range of differences between the meanings of each concept. A discussion of the reasons for this vast scope of different meanings is also conducted.

1. Introduction

A frequent criticism in the US has been the extensive use of multiple-choice tests, which has led to an upsurge in the interest in so-called alternative assessments in the US in the 1990s (Kirst, 1991; Messick, 1994). This growing interest has resulted in a more frequent use of these kinds of assessment (Herman, 1997) as well as in an extensive literature on the subject (Arter & Spandel, 1992). The body of literature on performance assessment and authentic assessment has been considerably enlarged (Hambleton & Murphy, 1992; Terwilliger, 1997). However, the literature manifests a considerable lack of agreement on the meanings of these terms.

---

[1] This report is also published as Research report No 4, 2001, in Mathematics Education, Department of Mathematics, Umeå University.

A big issue in a number of countries has been the experienced lack of realism in many contextualized tasks (e.g. Boaler, 1993; Cooper, 1992; Greer, 1993; Verschaffel, De Corte & Lasure, 1994). The concern is that these 'dressed up' tasks might have a negative impact on students' learning and attitudes, as well as on the construct validity of assessments including these tasks. The tasks are said to be non-engaging (Boaler, 1994; Gerofsky 1996; Sowder, 1989) and not enhancing the development of student beliefs that regard school mathematics as useful for situations in life beyond school (Boaler, 1994; Nesher, 1980). They are also said to promote superficial solution strategies not primarily based on the mathematical properties of the situation at hand (Greer, 1992; Reusser, 1988). Such strategies do not involve considerations of real life knowledge (e.g. Reusser & Stebler, 1997; Verschaffel & De Corte, 1997), and would not be suitable for problem solving in real life beyond school. They also seem to exhibit social class (Cooper & Dunne, 2000) and gender bias (Boaler, 1994). For reasons such as these a shift towards more 'realistic' or 'real life' tasks has been advocated. But just as with performance assessment and authentic assessment, the meanings of these terms are unclear in many publications, and vary substantially between authors.

Section 2 in this article deals with the meanings of *performance assessment*. A number of the meanings attributed to performance assessment focus on the response format. Different requirements of the response format discriminate between the different definitions, and the exclusion of multiple-choice format is a common factor among these meanings. Another category of meanings defines performance assessment as a relatively direct assessment, in the sense that there is a close similarity between the observed performance and the performance of interest, thus requiring the observed performance being more than a valid indicator of the performance of interest. Some of these definitions demand that students' work include non-written performance. *Authentic assessment* is treated in section 3. This concept is often associated with assessment emulating real life task situations, but also possesses meanings such as assessment aligned with curriculum and assessment that effectively support learning. Other meanings emphasize the construct

validity of assessment, irrespective of whether the assessment is, for example, aligned with the curriculum or not. In section 4 *realistic* and *real life tasks* together with similar terms are discussed. Realistic tasks emulate task situations in life, but different demands on the fidelity of the representations of the emulated task situations and different meanings assigned to the term 'realistic' cause differences in the definitions of realistic tasks. Real life tasks have also been defined as school tasks emulating real life task situations, but sometimes they are seen as the actual out-of-school task situations. Different demands on the task structure produce different meanings of the definitions based on these tasks being school tasks. All of the concepts treated in this paper possess a multitude of meanings depending on the author. In addition, in many publications they are not defined at all, which due to the many possible meanings sometimes makes the authors' descriptions and points difficult to interpret. The credibility of these concepts in education reform would be strengthened by clarity of the full scope of different meanings these concepts possess, and the explicit declaration of the intended meaning in each publication (c.f. Newmann, 1998; Terwilliger, 1997, 1998).

The aim of this article is neither to present additional definitions nor to make judgements on existing ones. The intention with the article is to describe different meanings of these concepts in an attempt to clarify their diversity as well as their similarities. Most definitions of performance assessment seem to be subject independent and therefore section 2 mostly deals with general definitions not specific for mathematics. Definitions of authentic assessment are also often general, but not to the same extent as for performance assessment. Both general and mathematics specific definitions will therefore be included. The first subsection on authentic assessment provides a classification of different meanings, and is followed by two subsections with examples of definitions. These two subsections are intended to clarify the classification with the inclusion of definitions not specific to mathematics and mathematics specific definitions in subsection 3.2 and 3.3 respectively. The section on authentic assessment concludes with a short summary. Since realistic and real life tasks do not seem to be discussed generally in the literature to the same extent as the first two concepts,

section 4 considers only mathematics specific definitions. The paper ends with a discussion of possible reasons for the extreme diversity of meanings attached to some of the concepts treated in this paper.

2. Performance assessment

The literature on the concept of performance assessment is extensive and the selection of references and the disposition have been made so that the broad spectrum of differences as well as similarities between different meanings will be as clear as possible. From the exposition it will be evident that, depending on the author, the concept of performance assessment can mean almost anything. It may even include multiple-choice tests!

Performance assessment is said by its advocators to be more in line with instruction than multiple-choice tests. With an emphasis on a closer similarity between observed performance and the actual criterion situations, it can also in a positive way guide instruction and student learning and promote desirable student attitudes. Furthermore, it is viewed as having better possibilities to measure complex skills and communication, which are considered important competencies and disciplinary knowledge needed in today's society.

In addressing the issue of the meaning of the concept of performance assessment it can be helpful to recognize that there is often a gap between the *characteristics* and the *definitions* of performance assessment outlined in the literature, although it is not always explicit. When performance assessment is described in terms of its characteristics, that is, by means of typical properties of such assessments, the descriptions mostly involve cognitive processes required by the students, but also the inclusion of contextualized tasks and judgmental marking in the assessment. Examples of phrases characterizing performance assessment are higher levels of cognitive complexity, communication, real world applications, instructionally meaningful tasks, significant commitments of student time and effort, and qualitative judgements in the marking process. When concrete examples are given, they are mostly in very close resemblance with criterion situations, demanding higher order thinking

and communication, or involving students in accomplishments with value beyond school, for example driving tests and making paintings. Furthermore, in most cases the characteristics describe the aims and possibilities of performance assessment and not its boundaries. Not surprisingly they reflect the goals said to be better assessed with performance assessment. The definitions of performance assessment put forth are of a different kind. When performance assessment is described by means of some kind of definition, in the sense that the description states a more precise meaning of the concept, then the boundaries are more noticeable. The definitions of performance assessment vary widely, both in focus and in possible interpretations of what is actually to be regarded as performance assessment. In the following a guided odyssey over different definitions is undertaken.

In the definition made by the Office of Technology Assessment, U.S. Congress (OTA, 1992), performance assessment is defined by means of response format. According to this definition all kinds of assessment, except those with multiple-choice response formats, are regarded as performance assessment.

> It is best understood as a continuum of formats that range from the simplest student-constructed response to comprehensive collections of large bodies of work over time.... Constructed-response questions require students to produce an answer to a question rather than to select from an array of possible answers (as multiple-choice items do) … examples include answers supplied by filling in the blank; solving a mathematics problem; writing short answers;…
> (Office of Technology Assessment, U.S. Congress, 1992, p. 19)

Arter (1999) also focuses on response format but demands more of performance assessment. Quoting Airasian (1991) and Stiggins (1997), she defines performance assessment as “assessment based on observation and judgement”. Arter points to her view of the relation to constructed response, which leads to a slight difference in assessment classification compared with the OTA: “Although fairly broad, this definition is not intended to include *all* constructed-response-type items (especially short answer and fill in the blank), but, admittedly, the line between constructed response and performance assessment is thin”.

Airasian (1994) implicitly addresses this difference between any constructed response and performance assessment. Performance assessment of intellectual abilities such as solving a mathematics task is said to demand insight into students' mental processes. According to Airasian this can be achieved when students have to show the work carried out to solve the task. This is, he claims, in contrast with most paper-and–pencil test items, where the teacher observes the *result* of the pupils' intellectual process but not the thinking that produced the result. When students are only required to show the end result of their work there is little direct evidence that the pupils have "followed the correct process" (Airasian, 1994, p. 229).

In Kane, Crooks & Cohen (1999) however, the definition of performance assessment does not have to do with response format. They claim that all assessments demand some kind of performance from the examinees and that choosing an alternative is also a performance. The performance required by the students is not enough to classify the assessment. It has to be seen in relation to the particular performance of interest. They visualize with an example:

> If we were interested in how well examinees can paint landscapes, an essay test on the theory of painting would probably not be considered performance assessment […] Note that the distinction being made here is not simply that between a test that is valid for an interpretation and one that is not. Rather, it is a distinction between a direct measure and an indirect measure or between a sign and a sample. Even if an essay test on painting theory had a near perfect correlation with skill in painting landscapes (as it might in a population in which some people had studied painting and some had not) and therefore constituted a valid indicator of skill (at least in this population), it would probably not be considered a performance test because it does not involve the performance of interest, which includes putting paint on canvas. What then, is special about the performances required in performance assessment? It seems that the defining characteristic of a performance assessment is the close similarity between the type of performance that is actually observed and the type of performance that is of interest.
>
> (p. 6-7)

This approach, emphasizing simulation instead of response format in defining performance assessment, is also adopted by other authors with somewhat different emphasis. Shepard and Bliem (1995) specifies the performance of interest as “the actual tasks and end performances that are the goals of instruction” (Shepard & Bliem, 1995, p. 25), and in the definition in the Glossary of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999), the performance of interest is explicitly connected to performance in ‘real life’:

> **performance assessments** Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.
>
> (p. 179)

A conceptually different approach is adopted by Berk (1986). According to his definition a single event cannot be regarded as a performance assessment. A variety of instruments and strategies must be used on a number of occasions to collect data for the purpose of making decisions on individuals. Furthermore, the focus must be on systematic observations of non-written performances. However, this does not mean that the arsenal of usable measurement instruments in performance assessments cannot include tests focusing on paper-and-pencil written responses. In fact, even multiple-choice tests may be used according to this definition of performance assessment. (According to Berk, a test that is used on a single occasion can be a performance *test*. In such a test the performance of interest “is demonstrated through directly observable behavior as opposed to paper-and-pencil written response” (Berk, 1986 p. ix)).

The concept of performance assessment as it is used in the TIMSS study (Harmon et al., 1997), also requires some sort of practical activity. The students are provided with instruments and equipment as a means to create an environment that is considered to be more like situations encountered in life beyond school than those offered by traditional paper-and-pencil tests. There is, however, a fundamental difference between

this definition and the definition proposed by Berk (1986). In the definition by Berk the observation is intended to be direct, in the sense that the observed performance is the performance of interest. In the TIMSS definition the observed performance does not necessarily have to resemble the performance to which inferences are made. The instruments and equipment are provided merely as a means to elicit performance that is a more valid indicator "of the students' understanding of concepts and potential performance in real life situations" (Harmon et al., 1997 p. 5) than the performance measured by means of traditional paper-and-pencil tests.

All the mentioned definitions are general in the sense that they are subject independent. Often when performance and performance assessment are discussed, even for specific subjects such as mathematics, a general definition is called upon. In the NCTM Assessment Standards for School Mathematics (National Council of Teachers of Mathematics, 1995) performance is defined as:

> The carrying out or bringing to completion of a physical activity or production of some significance, which displays one's knowledge and judgment while engaged in the task.
>
> (p. 90)

However, subject-dependent definitions can also be found in the literature. For example, Solano-Flores & Shavelson (1997) have presented a definition specifically for science. In this definition they demand both well-contextualized tasks and hands-on activities. They also require a response format and a scoring system that facilitate the assessment of students' scientific reasonableness and accuracy. They claim that there is a knowledge domain associated with what has been lumped together as "science process skills" and that "SPA [science performance assessment] can be conceived as tasks that recreate the conditions in which scientists work and elicit the kind of thinking and reasoning used by scientists when they solve problems" (p. 18). Thus, they state the kind of knowledge and skills required of the students as well as the material involved.

In summary, most definitions offered for performance assessment can be viewed as *response-centered* or *simulation-centered*. The response-centered definitions focus on the response format in the assessment, and the simulation-centered definitions focus on the observed student performance, requiring that it is similar to the type of performance that is of interest. In some of the simulation-centered definitions practical activity, through the use of equipment not normally available on paper-and-pencil tests, are required. There are substantial differences between definitions belonging to the different categories. For example, many assessments that would be regarded as performance assessment by the definition of the OTA (1992) would not be considered to be performance assessment with the requirements of Kane et al. (1999). The requirements by the OTA that assessments built up by tasks with any response format requiring student-constructed response (such as filling in the blank) are performance assessments are significantly different from the requirements by Kane et al. (1999) that the observed student performance must be similar to the type of performance of interest. There are also significant differences between the definitions within each category. Within the response-centered category different definitions can be placed on a continuum of different strength of the demands on the responses. On the one end of this continuum there is the definition by the OTA, which displays a marked difference from, for example, the definition by Airasian (1994) that requires the thinking that produced the answers to the tasks to be explicitly shown. Considering the equipment requirements in some of the simulation-centered definitions, it is clear that there are also significant differences within this category. In addition, acknowledging the relative aspect of the broad simulation-centered definitions, there are most certainly also significant problems in the interpretations of these definitions. The focus on high fidelity simulations can, for example, be interpreted as a requirement for assignments taken directly from real life experience, with no other limits in the examinee's access of tools, collaboration, and literature etc, than in the simulated real situation. It can also be interpreted as an assessment administered for classroom use, demanding only the traditional mathematics word problems requiring short student-constructed responses. Fitzpatrick &

Morrison (1971) expressed this relative aspect of this approach of classifying assessments:

> It is apparent that there are many degrees and kinds of artificialities in tests. A test of the class here designated as *performance and product evaluation* is one in which some criterion situation is simulated to a much greater degree than is represented by the usual paper-and-pencil test. This type of test usually is called a performance test, and that term will be used here interchangeably with the more complete *performance and product evaluation* […] There is no absolute distinction between *performance tests* and other classes of tests - the performance test is one that is *relatively* realistic.
>
> (p. 238)

Thus, analyzing different definitions, it is clear that some of them share important properties. At the same time it is also evident that performance assessment can mean almost anything. Furthermore, returning to the issue of a gap between the characteristics and definitions of performance assessment, it is clear that performance assessment by most definitions demand only very few of the characteristics mentioned earlier in this paper. Tasks need not for example be real world applications or require much communication and high levels of cognitive complexity just because students' activities are hands-on or because they have to construct an answer themselves. Since it takes both knowledge and a considerable amount of time to construct high quality assessments, it is likely that a large number of assessments called performance assessment lack many of the characteristics attributed to it in the literature. It is obvious that student-constructed response (beyond selecting from a set of ready-made answers) is a prerequisite for students' extended communication. It is likely that such tasks can be experienced as instructionally more meaningful than multiple-choice tasks. It is also possible that instruments and equipment have the possibility to elicit performance that is a valid indicator of performance in real life situations. However, there is not a one-to-one correspondence between the frames of students' task solving and the performances and experiences sought after. It may therefore be useful to be clear about whether it is the typical properties, aims or the definition of performance assessment that is discussed in a publication.

3. Authentic assessment

As in the case with performance assessment authentic assessment can also mean almost anything. Subsection 3.1 includes a description of perspectives and foci on authenticity in assessment. The description outlines major directions of different kinds of meanings attributed to authentic assessment, and can serve as a classification of the various meanings of the concept. In subsections 3.2 and 3.3 the different perspectives and foci are exemplified through a number of definitions of authenticity. Subsection 3.2 deals with general definitions and subsection 3.3 with definitions in the special case of school mathematics. Some concluding comments are made in a short summary in subsection 3.4. There is also extensive literature on authentic assessment. The ambition has been to select illustrative examples of the perspectives and foci. Thus, the definitions included are intended to exemplify and clarify the perspectives and foci, outlining their consequences in the form of differences as well as similarities between the meanings of the concept of authentic assessment. The aim is not to capture every aspect of the different meanings in detail but to outline fundamental features that have been identified. The main focus of this section is on authentic assessment. However, since tasks play a central role in many assessment forms, and since they have to be regarded as authentic for such assessments to be authentic, ideas focusing on tasks are considered as well.

3.1 Classification of meanings

The concept of authentic assessment is a much more recent term than performance assessment even if some authors considers it equivalent to performance assessment. It originated from dissatisfaction with standardized norm-referenced tests and the consequences of that assessment practice. In the Oxford Dictionary (1988) a*uthentic* is explained as “genuine; known to be true” and in the Cambridge Dictionary (1996) as “real or true; being in fact what it is claimed to be; GENUINE”. In relation to assessment, the explanations in the dictionaries can be interpreted as what is claimed in or by the task or assessment is really true. The fact that something is supposed to be true, however, gives the concept different meanings depending on the chosen

frame of reference. The meaning of the word authentic makes the choice of focus an open question, and different foci have also been applied in the literature. Two main questions are of interest here: *what* it is that is supposed to be real or true, and what it is that it is supposed to be *true to*? Three main *perspectives* have been used in relation to the second question:

- *Life beyond school*. With this perspective similarity to life beyond school is emphasized. This can include the requirement that students are engaged in cognitive processes that are important for successful adult accomplishments, the requirement that students are working with tasks that are of importance in life outside school, or the requirement that students are engaged in assessments under the same working conditions (e.g. time constraints and access to relevant tools) as they would have had in life beyond school.
- *Curriculum and classroom practice*. In this perspective the authenticity lies in the resemblance to the curriculum or to classroom practice. Examples of important assessment features in this perspective are curriculum alignment and concordance in students' working conditions during assessment and classroom practice.
- *Learning and instruction*. This perspective is based on the idea that an important purpose of assessment is learning. Assessments are authentic if they are effective for learning or for guiding instruction. Such assessment could involve self-assessment or tasks designed to provide information that is useful for guiding further learning and instruction. The emphasis on the formative aspect of assessment is a main difference between this and the other two perspectives.

In relation to the first question "*what* it is that is supposed to be real or true" three main *foci* have been identified:

1. *Processes and products*. This focus deals with cognitive processes, performances, constructs, or products that students engage in, produce, or are assessed on. Some authors have specific processes or products in mind that are claimed to be important, and some others are more unspecific about these processes or products. In both cases the processes or products are regarded as the important issue in authenticity. The task, activity or assessment is regarded as authentic

if, for example, students are engaged in cognitive processes that are important in successful adult behavior in life beyond school, meet curricula goals, or are effective in the learning process.

2. *Conditions*. With this focus authenticity is dependent on the conditions, under which the student activity takes place, being true to some main perspective above. This could mean, for example, that time constraints and access to relevant tools are the same in the assessment situation as in some situation in life beyond school or in ordinary classroom practice. The third perspective, learning and instruction, requires that assessment procedures promote a situation that is effective for learning (this could, for example, mean that student involvement in all phases of the assessment is required).
3. *Figurative context*. Here the focus is on the *figurative context*, that is, the situation described in the task (Clarke & Helme, 1998). The figurative context has to be faithful to some subject or field of application outside mathematics. Authenticity lies in the figurative context consisting of problems and objects actually belonging to that field, for example a potential task situation in physics studies or in life beyond school capturing the important contextual aspects of that situation.

This does not mean that the perspectives or foci are totally independent of each other, nor that the authors are only interested in one perspective and one focus. It merely means that these perspectives and foci represent different frames of reference chosen in defining authenticity, resulting in different meanings of the concept.

In addition to the perspectives and foci mentioned some authors do not emphasize any perspective. Instead they put emphasis on procedures of assessment or on quality features of assessment, such as construct representation (the match in important aspects of the underlying theoretical construct intended to be measured and the aspects of the construct and other factors captured by the assessment), that are more or less independent of the described perspectives.

3.2 Examples of definitions

According to Cumming & Maxwell (1999, p. 178) and a discussion in *Educational Researcher* (e.g. Terwilliger, 1998) the first formal use of the term 'authentic' in the context of learning and assessment appears to have been made by Archbald & Newmann (1988). Archbald & Newmann acknowledged that "traditional tests" have been criticized for neglecting the kind of competence needed for dealing successfully with many situations beyond school. They stated that assessment should not measure just any kind of achievement, but valuable or meaningful forms of mastery. These forms of mastery are the intellectual qualities they considered to be needed for many significant human accomplishments. Newmann describes authenticity as a key facet of intellectual quality defined as:

> the extent to which a lesson, assessment task, or sample of student performance represents construction of knowledge through the use of disciplined inquiry that has some value or meaning beyond success in school.
>
> Newmann (1997, p. 361)

That is, students should construct knowledge. But even though construction of knowledge is based on the understanding of prior knowledge, reproduction of knowledge that others have produced does not suffice for authentic academic achievement. The students also have to relate to that knowledge through interpretation, evaluation, synthesis, or organization of information. The cognitive work that has to be applied is disciplined inquiry. Students should engage in the use of prior knowledge to get beyond that knowledge, establish relationships between pieces of this knowledge to construct in-depth understanding around a reasonably focused topic, and conduct their work and express their conclusions through elaborate communication. Authentic achievement is also said to have "*aesthetic*, *utilitarian*, or *personal value* apart from documenting the competence of the learner" (Newmann, 1997, p. 365). The students might be faced with tasks that are similar to what they have encountered or are likely to encounter in life beyond school and they might be requested to present their work to an audience beyond school.

In this view of authenticity the main focus is on students' cognitive processes and products. The desired cognitive processes are in the form of disciplined inquiry and the desired product is the production of knowledge beyond the mere reproduction of presented knowledge. The third criterion of authenticity, that the accomplishment should have value beyond school, is also related to the desired product but could also be seen as requiring the figurative context dealing with issues that have meaning beyond school. However, to be engaged in authentic adult work, Newmann & Archbald (1992) argue, there must be at least four essential working conditions. The students must sometimes have the opportunity to collaborate with others, they must have access to tools and other resources, they must to some extent have the possibility to influence the conception, execution, and evaluation of their work and they must have flexible use of time. They also state the importance of the assessments having criterion-based standards, multiple indicators and including human judgement. They claim that these conditions must be fulfilled in assessment of the human accomplishments described by the concept of authenticity. Thus, the defining features of authenticity are the specific processes and products considered important in the perspective of life beyond school. In addition, some conditions, also in relation to life beyond school, are considered to be necessary for authenticity.

Brown, Collins & Duguid (1989) also focus mainly on 'processes and products' with the perspective of 'life beyond school'. However, while Archbald & Newmann refer to specific and identified processes and products considered to be shared by all significant adult accomplishments depending on formal knowledge, Brown et al. regard "authentic activities" to be related to practitioners in different fields, possibly different activities in different fields. Thus, the unspecified "activities" of Brown et al. may differ from those emphasized by Archbald & Newmann. The following two quotations serve to illustrate this point, the second citation indicating that mathematics can be a culture of its own, possibly different from the culture of, for example, historians.

Authentic activities then, are most simply defined as the ordinary practices of the culture. […] Classroom activity very much takes place within the culture of

schools, although it is attributed to the culture of readers, writers, mathematicians, historians, economists, geographers, and so forth.

Brown, Collins & Duguid (1989, p. 34)

By allowing students to generate their own solution paths, it helps make them conscious, creative members of the culture of problem-solving mathematicians.

Brown, Collins & Duguid (1989, p. 38)

Similar ideas of authenticity as Brown et al., but specifically in relation to assessment, are presented by Wiggins. His perspective of authentic assessment is 'life beyond school' and in addition to 'processes and products' he also emphasizes 'conditions'. He claims that in authentic assessment "The tasks are either replicas of or analogous to the kinds of problems faced by adult citizens and consumers or professionals in the field" (1993, p. 206), and that "replicating or simulating the diverse and rich contexts of performance" (1993, p. 207) is the most important one of his nine criteria of authenticity. This rich context of performance is partly provided by the conditions of the assessment (e.g. time constraints) and partly by the figurative context. However, the acceptance of analogous kinds of problems leaves out an essential part of a definition focusing on the figurative context (see e.g. Niss (1992) below).

Shepard (as quoted by Kirst, 1991) takes another approach to authentic assessment. She gives the concept of authentic assessment as a synonym to performance assessment.

Use of the term *authentic assessment* is intended to convey that the assessment tasks themselves are real instances of extended criterion performance, rather than proxies or estimators of actual learning goals. Other synonyms are *direct* or *performance assessments*.

(Kirst, 1991, p. 21)

Not only does this view put higher demands on the similarity between the type of performance that is actually observed and the type of performance of interest than Wiggins does (who considers *analogous* kinds of problems to those of interest to be sufficient), but it is also conceptually different from the intentions of e.g. Archbald & Newmann (1988), and Wiggins (1989). While the emphasis of Archbald &

Newmann and Wiggins is on the alignment between assessment and, by the researchers, stated and desired learning goals, Shepard is concerned with the alignment between assessment and *any* actual learning goal. Like Shepard, Messick (1994) focuses on perspective-independent quality features of assessment, and is not concerned with any particular 'perspective'; "…I must at some point broach the questions of authentic to what criterion and for what purpose? I then argue that other forms of assessment are more constructively characterized not as inauthentic, as the rhetorical innuendo implies, but as authentic to other criteria and purposes" (p. 14). What is at the heart of the matter is the construct validity of the assessment of "complex of knowledge, skills or other attributes that are tied to the objectives of instruction or otherwise valued by society" (p. 16). However, Shepard's definition seems to imply that authentic assessment requires that both major threats to construct validity, construct underrepresentation and construct-irrelevant variance, are minimized. That is, for appropriate interpretations of assessment results the complexities of the underlying theoretical construct must be captured in the assessment, while irrelevant factors must not be, and that is required of an authentic assessment. Messick, on the other hand, defines authenticity in assessment as only minimal construct underrepresentation (and regards construct-irrelevant variance as the implicit validity standard for *directness* of assessment):

> The basic point in this discussion of complex and component skills is that the validity standard implicit in the concept of authenticity appears to be the familiar one of construct representation […]. That is, evidence should be sought that the presumed sources of task complexity are indeed reflected in task performance and the complex skill is captured in the test scores with minimal construct underrepresentation.
>
> Messick (1994, p. 20).

Hughes (1993) focuses on 'processes and products' and 'conditions', and has both 'life beyond school' and 'curriculum and classroom practice' as her main perspectives. Together these perspectives and foci constitute the basis for her definition of authentic assessment. An assessment measure is said to be authentic if "it faithfully reflects 1) the knowledge and process that form the basis of the subject matter and 2) the conditions

under which the achievement normally takes place". The base for authentic measures are activities that reflect "tasks typical of classrooms and real-life settings" (p. 29).

Shifting the main perspective to learning and instruction changes the meaning of authentic assessment. Terry & Pantle (1994) claim that the purpose of authentic writing assessment is to support literacy learning and that assessment must therefore meet and support the components of learning in order to be authentic. They argue that this is the case when students' activities have meaning beyond assessment, when assessments occur over time and contexts, when they involve learners through, for example, self-evaluation and collaboration with teachers and peers, and when they allow students to feel comfortable and confident in their writing, for example through evaluation based on students' strengths. Thus, the main focus is on 'conditions' that are required to underpin an environment that supports learning. These requirements are partly different from those required by the authors mentioned above. For instance, the demand that students feel comfortable and confident during assessment is a significantly different requirement than any of the requirements mentioned above.

In addition to definitions based on essentially one main perspective and one or two foci, there are definitions based on several perspectives and foci and on perspective-independent assessment quality features. They include a mixture of features mentioned in this paper and other features that can be traced back to the perspectives and foci presented here. Examples of the latter are the requirements that authentic assessments "give students both feedback upon completion of the project" as well as "guide their work along the way" (Schack 1994, p. 39), include multiple measures (including observers, that is "being there"), and minimize biases (Maker, 1994).

Finally, a description by Baker & O'Neil (1994) of authenticity in assessment calls our attention to another important issue of authenticity, namely authentic to whom? Baker & O'Neil claim that authenticity in assessment lies in the tasks being contextualized and "intended to be

inherently valuable to students, either immediately or because they can see its longer-term connection to an important goal" (p. 15). The word 'intended' in the description by Baker & O'Neil reveals a focus on the *assessment constructor's* view of the task or assessment. The consequence of such a formulation, concerning authenticity, is that it does not primarily matter whether we succeed in constructing tasks that are experienced as valuable by the students. Another issue of the description by Baker & O'Neil is the time perspective of the term 'valuable'. The possibility of the tasks being valuable to students if they spot a longer-term connection to an important goal makes it possible for almost any task to have this authenticity aspect. Some students might very well experience some basic algorithms as being 'longer-term connected' to an important goal, whatever the context. Such a broad meaning of the term 'value' seems quite different from the description of authentic tasks involving the term 'value' by Newmann (1997, p. 366-367). He requires authentic tasks to "result in discourse, products, and performance that have value or meaning beyond school", and states 'problems connected to the world' and 'audience beyond school' as the two task standards that reflect 'value beyond school', which is a more general standard for authentic achievement. These issues may at first glance be seen as trifling technicalities but may prove to be crucial in constructing, evaluating and revising assessments as well as for the meaning of the concept of authentic assessment.

3.3 Examples of definitions specific to school mathematics

In defining authenticity in the special case of school mathematics some authors call upon a general definition. Other authors add mathematics specific meanings to a general definition or states an entirely mathematics specific definition. The NCTM Assessment Standards for School Mathematics (National Council of Teachers of Mathematics, 1995, p. 87) refer to Wiggins for their definition of authenticity, without any further subject specific addition. The Cognition and Technology Group at Vanderbilt, CTGV, (1990) adopted the general definition of authentic activities by Brown et al. (1989, p. 6), and for mathematics tasks they specified two levels of authenticity, both involving the figurative context. One level involves the authenticity of the objects and

data in the setting shown in a video, and the other level involves the authenticity of the assignment of the task. Both levels are seen in the perspective of life beyond school. Stenmark (1991) addresses the issue of the influence of the activities within the specific discipline of mathematics on a definition of authenticity. She describes an authentic assessment task in general terms as: “The task uses processes appropriate to the discipline” and “students value the outcome of the task” (p. 16), and clarifies the mathematics specificity as:

> They involve finding patterns, checking generalizations, making models, arguing, simplifying, and extending-processes that resemble the activities of mathematicians or the application of mathematics to everyday life.
>
> (Stenmark 1991, p. 3)

Also Romberg (1995) offers a clarification of the activities of the discipline in his definition of authentic assessment in mathematics. He claims that authentic assessment of student performance are assessments that are “trustworthy indicators of mathematical power”, mathematical power being defined as “an individual’s abilities to explore, conjecture, and reason logically, as well as the ability to use a variety of mathematical methods effectively to solve nonroutine problems” (Romberg 1995, p.vii). According to Lesh & Lamon authentic mathematical activities require the use of “ a representative sample of the knowledge and abilities that reflect targeted levels of competence in the field” (1992, p. 17). While Romberg uses the terminology of NCTM, Lesh & Lamon refer to the Mathematical Sciences Education Board for the specific nature of mathematics that defines authentic mathematical activities. MSEB describes the discipline of mathematics in words of “the science and language of patterns” (MSEB 1990, in Lesh & Lamon 1992 p. 24), which however is concluded by Lesh & Lamon, referring to Earnest (1991), to be in consistency with the view of NCTM. In addition to this focus on ‘processes and products’, Lesh & Lamon also focus on ‘conditions’ and ‘figurative context’ in their requirements of authentic mathematical activities, all seen in relation to life beyond school.

An influence of the specific nature of mathematics is also present in an attempt by Lajoie (1995) to define some tentative principles for an operational definition of authentic assessment to improve learning in the area of school mathematics. These principles involve the requirement of alignment with the "NCTM *Standards*". However, in addition to the cognitive dimensions she also proposes that information should be gathered on conative dimensions (e.g. students' interests, perseverance and beliefs) recognized to affect learning. The perspectives of 'learning and instruction', and 'curriculum' ("NCTM *Standards*"), as well as involvement of the issues of assessment procedures and assessment quality features that are independent of the perspectives also lead to additional requirements that are similar to those presented in subsection 3.2.

The focus on the figurative context is also represented in mathematics specific definitions of authenticity. The definition by Niss (1992) and the definition in the mathematical literacy framework of the OECD's Programme for International Student Assessment, PISA, (Organisation for Economic Co-operation and Development, 1999) are examples of definitions in which the focus is on the figurative context and neither 'processes and products' nor conditions are the main issues. The issue is that the figurative context truthfully describes a situation from real life that has occurred or might very well happen. In these definitions also other school subjects than mathematics, in addition to life beyond school, is accepted as an arena for the situation. Niss defines such situations as:

> We define an *authentic* extra-mathematical situation as one which is embedded in a true existing practice or subject area outside mathematics, and which deals with objects, phenomena, issues, or problems that are genuine to that area and are recognized as such by people working in it
>
> Niss (1992, p. 353)

Thus, for a task to be authentic the figurative context has to include real components and realistic data, and someone in the described setting must be likely to be called upon to address the problem as it is presented in the task. Since it is possible to be engaged in the activities of a discipline without the figurative context being true to a specific situation, this is a

different definition of authentic tasks than the definitions offered by, for example, Brown et al. (1989) and CTGV (1990).

3.4 Summary

It seems that similarities between different definitions of authentic assessment often reflect the same choices of perspectives and foci, even if shared features can also be found in definitions where different perspectives and foci can be recognized. However, the identified perspectives and foci also visualize prominent differences in the meanings of authenticity. Definitions of authentic assessment display such differences as requiring the assessment of specific cognitive processes and products (Archbald & Newmann, (1988), being synonymous with assessments by which the assessed skills are captured with minimal construct underrepresentation (Messick, 1994), and requiring the assessment to be formative as well as supportive of the students' possibilities of feeling comfortable (Terry & Pantle, 1994). In addition, the descriptions of authenticity and authentic assessment are often quite indistinct and sometimes even contradictory within the same publication.

4. Realistic and real life tasks

In this section the concepts of *realistic*- and *real life tasks* (and similar terms such as *real world problems*) are treated. Although the meanings of these concepts vary, they are less disparate than those of performance and authentic assessment. Since their meanings are also less clarified in the literature they will be more briefly dealt with. When the notion of 'problem' is an issue, caution is in place. Sometimes 'problem' and 'task' are considered equivalent in meaning, but at other times the term 'problem' is restricted to non-routine tasks. However, when used together with the terms 'realistic', 'real life' and 'real world', the possible difference in meaning between these two concepts has rarely been an issue. Consequently, it has not been considered meaningful to treat them as different concepts in this exposition.

The word *realism* is described in the Oxford Advanced learner's dictionary (1988) as "showing of real life, facts, etc in a true way,

omitting nothing that is ugly or painful, and idealizing nothing". It is in this sense that the word is used as an attribute to tasks and problems in the education literature. However, different demands on the degree of fidelity of the representation of real life and different views of the concept of reality result in different meanings of the notion. These differences can be exemplified by referring to Van den Heuvel-Panhuizen (1996), Burton (1993) and Cooper & Dunne (2000). Van den Heuvel-Panhuizen describes the instructional approach *Realistic Mathematics Education* in the Netherlands, in which the link to reality is of great significance. However, the distinctive feature of tasks included in RME is not fidelity to situations in actual life, but fidelity to situations that can be imagined by students in such a way that they take an active interest in the activity of solving the task. Thus, in this case realism seems to be interpreted as what could be described as imagined realism:

> The contexts need not necessarily refer, however, to real life situations. The important point is that they can be organized mathematically and that the students can place themselves within them.
>
> (Van den Heuvel-Panhuizen, 1996, p. 13)

Burton (1993), on the other hand, restricts the notion of realistic problems to those dealing with situations encountered in real life:

> Why are puzzles, games, riddles, logical challenges so motivating for young people? Precisely because of the enjoyment that they gain from using the mind to tackle something which is new to them even if it is not 'realistic' in the sense of embedded in their life-style concerns. At the same time, realistic problems are equally important to ensuring that learners perceive that mathematics does contribute to working at and resolving issues of living.
>
> (Burton, 1993, pp. 11-12)

Cooper & Dunne (2000) also require that aspects of real life are present in realistic tasks. However, the demand that a realistic task "contains either persons or non-mathematical objects from 'everyday' settings" (p. 84) is not a very strong condition. In most cases this does not suffice for students to perceive the power of mathematics in dealing with issues of

real life, which is an aim expressed by Burton for tasks she labels realistic.

Concepts similar to ‘realistic tasks’ are *real life-* or *real world tasks* (or *problems*). It seems that the meanings of these concepts can be described in relation to two dimensions, the *situation* in which the task is encountered and the *task structure*. The situation is a dichotomous variable. Either real life tasks are seen as tasks encountered in real life beyond school (e.g. Illingworth, 1996), or they are conceived of as school tasks similar to tasks encountered in life beyond school (e.g. Swan, 1993). Both of these kinds of tasks often require mathematization and the use of both mathematical and non-mathematical knowledge in order to gain new insights into the real world. However, the difference between a school task and a corresponding task situation in real life beyond school may be huge and the affect this difference has on students’ task solving may also be significant. Partly due to the influence the context in which the task is encountered have on students’ task solving, but also because simulating real life situations is a complex business and different simulations of the same situation can be carried out with very different results, that is, resulting in very different school tasks. The dimension of task structure is either not addressed at all (e.g. Boaler, 1994), or is discussed in detail as a salient dimension of the nature of real life problems (e.g. Borasi, 1986; Illingworth, 1996). Borasi (seeing real life problems as tasks in education) identified four structural, constitutive elements of problems (or related concepts such as tasks). She uses the term *real-life problem* to describe situations where the *context* and problem *formulation* have not been fully provided in the text, and the problem solver will need to find extra information and refine the formulation of what he or she is expected to do. In solving the problems there are many possible, and only approximate *solutions*, and the *methods of approach* involve exploration of the context, reformulation and the creation of a model. Illingworth, on the other hand, describes features of problems encountered in life beyond school. These features are partly similar to those proposed by Borasi. These problems are proposed to be complex, and require more than one step with different algorithms for a successful solution. The data given is often not the same

as that needed and in many cases several solutions are possible, which may result in awkward answers.

As stated above, performance assessment and authentic assessment are concepts with very disparate meanings, and they are too often used in the literature without clarification of their intended meanings. However, the terms realistic- and real life tasks (and similar terms) are even more rarely defined in texts. These terms may be conceived as self-explanatory but the above exposition describes a variety of meanings. Consequently, when the intended meanings of these concepts are not clarified in a publication the risk for the contribution to be obscure is significant.

5. Discussion

This paper is a part of a more extensive project investigating both the features of mathematics school tasks including an out-of-school figurative context and the impact of their concordance with real life task situations on student performance. In Palm (2001) a set of aspects of task situations in real life beyond school is proposed to be of importance in relation to those meanings of the concepts discussed in this paper that emphasize the similarity between tasks encountered in and out of school. The simulation of these aspects with reasonable fidelity is argued to facilitate the similarity between performance on school tasks and performance beyond school.

The analysis of the literature presented in this paper shows that the frequently addressed concepts of performance assessment and authentic assessment share some of the meanings given to them. Several of these meanings attributed to both concepts emphasize the use of tasks eliciting skills of important end goals of education by closely emulating task situations encountered in real life beyond school. However, several of the definitions of performance assessment provided in the literature emphasize response format and requirements of hands-on activities, features not prominent in definitions of authentic assessment. The definitions of authentic assessment, on the other hand, include meanings focusing on more or less specified cognitive processes argued to be important in life beyond school, and meanings requiring the figurative

context to be true to situations outside mathematics. Such properties are rarely the main issue for definitions of performance assessment. The most striking result of the analysis is however, the extent to which each of these concepts possesses different meanings. As described in this paper, these terms can mean almost anything. It is not unusual that concepts are not very well-defined and that they can possess slightly different meanings, but the concepts of performance assessment and authentic assessment have been given so many different meanings that the terms themselves practically no longer possess any meaning at all, although they are frequently used in the literature as if they had a well-defined meaning.

An explanation for this awkward state of the art may be found in the history of these concepts. At the middle of the 20th century the term performance test was in most cases connected to the meaning of practical tests not requiring written abilities. In education the idea was to measure individuals' proficiency in certain task situations of interest. It was acknowledged that the correlation between facts and knowledge, on the one hand, and performance based on these facts and knowledge, on the other, were not always highly correlated. Judgement of the performance in the actual situation of interest was therefore desirable. The usefulness of such tests was regarded as obvious in vocational curricula and they seem to have been mostly applied in practical areas such as engineering, typewriting and music. Out of school, such practical performance tests were for example used for considering job appliances and in the training of soldiers during the Second World War. In psychology, performance tests were mostly associated with non-verbal tests measuring the aptitude of people with language deficiencies (Ryans & Frederiksen, 1951). This historical heritage is still fundamental to the concept of performance assessment but now, at the turn of the century, the situation has grown considerably more complex. From the 1980s onwards there has been an upsurge in the amount of articles on performance assessment (the term assessment now coexisting with the term test). But now theoretical school subjects, such as mathematics, have also become a matter of interest. It is appropriate, at this point, to acknowledge the difference between vocational school subjects and theoretical school subjects, such

as mathematics as an independent subject, in terms of performance. In vocational subjects there are well-defined performances tied to the profession, which can be observed relatively direct ("the proof of the pudding is in the eating"). This is not the case for mathematics. Both a professional mathematician and a student may apply problem-solving techniques, but they solve very different problems and hence their performances are different. Students may occasionally be placed in task situations in real life beyond school so performance in such situations may be assessed relatively direct, but there is no well-defined performance tied to the understanding of mathematical concepts and ideas so inferences to such understanding can only be drawn from indicators. The growing interest in performance assessment and the new focus on more theoretical subjects seem to emanate from dissatisfaction with the extensive use of multiple choice tests in the US. The validity of these tests as indicators of complex performance was experienced to be too low, and to have negative effects on teaching and learning (Kane et. al, 1999; Kirst, 1991). When arguing for other forms of assessment better fulfilling these requirements the term performance assessment was recognized as a suitable choice. But desires for change open up numbers of possible perspectives, so new views on the meaning of the attribute 'performance' have been added, and demonstrably consensus on the meaning of performance assessment has not been reached. This variety of different meanings creates a great deal of confusion among people coming in contact with the term. However, coming to grips with the concept is probably even more difficult for people outside the US. Since most other countries do not have, nor have had, such an extensive use of multiple-choice testing there have not been similar requirements for change and in many countries a corresponding concept does not even exist.

The dissatisfaction with the emphasis on multiple-choice testing in the US was also a fundamental factor for the development of the concept of authentic assessment. This much more recent term in education arose from the urge to meet needs that were experienced not to be met by the use of multiple-choice tests. Norm-referenced standardized multiple-choice tests of intellectual achievement were said not to measure

important competence needed in life beyond school. Interpretations of test results from such tests were claimed to be invalid indicators of genuine intellectual achievement and since assessments influence teaching and learning they were also said to be directly harmful (Archbald & Newmann, 1988; Wiggins, 1989). From the original idea of assessing the important achievement defined by Archbald & Newmann (1988), a number of more or less related meanings have been attached to this concept. As for performance assessment, different purposes of reform and different views on for example knowledge, learning and assessment have probably contributed to the diversity of meanings. Linn, Baker & Dunbar (1991) identified three main arguments for reform in large-scale assessments in the US, namely the limitations of multiple-choice tests as tools for measuring towards "long-range objectives", for gathering clear and accountable evidence of what the students actually know, and as a participant in the possible use of the advances in cognitive and developmental psychology for linking assessment to theories of how learning occurs (p. 15-16). The accountability argument seems to imply a stronger demand for demonstration of the actual competencies desired, following that these arguments are consistent with the perspective-independent quality features of assessment and the 'perspectives' on authentic assessment identified in this paper. The choice of term (authentic) may perhaps also have added to the difficulties of maintaining a reasonably well-defined meaning. The term invites different foci and perspectives at the same time that it is extremely value laden – no one wants to construct an inauthentic assessment. The implication that everything else is inauthentic is contested by several authors (e.g. Messick, 1994; Terwilliger, 1997).

Speculatively, the diversity of meanings of the concepts of realistic and real life problems (and similar terms) is only a consequence of a natural distribution of meanings for concepts for which there have not been strong demands for clear definitions. Real life and realism can be seen in different ways, and this is reflected in the different meanings given to these concepts.

However, in the light of the various views on all of the concepts discussed in this paper it seems important to present a clear definition when addressing issues including these concepts. Due to the possibly vast differences between the simulation-centered definitions of performance assessment, a visualization of such definitions with non-obvious examples would many times be valuable as well. Given the difference in assessment practice (and history of assessment practice) around the world a clarification of the concepts seems particularly important when the publication is aimed at an international audience. When authors use these notions without specifying the intended meanings, it sometimes makes the contribution unclear. Except for potential misinterpretations of the contribution, this makes it difficult to scrutinize and criticize the contribution, which is a serious problem since this is a significant part of the enterprise of the research community.

Acknowledgement

The author wishes to thank Hans Wallin, Johan Lithner, and Mogens Niss for their most valuable comments to earlier versions of this paper.

References

Airasian, P.W. (1994). *Classroom Assessment* (2nd edition). New York: McGraw-Hill.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing. Washington*. DC: American Educational Research Association.

Archbald, D.A. & Newmann, F.M. (1988). *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School.* (Reston, VA, National Association of Secondary School Principals).

Arter, J. (1999). Teaching About Performance Assessment. *Educational Measurement: Issues and Practice, 18*(2), 30-44.

Arter, J.A. & Spandel, V. (1992). An NCME Instructional Module on Using Portfolios of Student Work in Instruction and Assessment. *Educational Measurement: Issues and Practice, 11*(1), 36-44.

Baker, E.L. & O'Neil Jr, H.F. (1994). Performance Assessment and Equity: a View from the USA. *Assessment in Education, 1*(1), 11-26.

Berk, R.A. (1986). Preface. In R.A. Berk (Ed.), *Performance Assessment: Methods & Applications* (pp. ix-xiv). John Hopkins University Press: Baltimore, Maryland.

Boaler, J. (1993). Encouraging the Transfer of 'School' Mathematics to the 'Real World' Through the Integration of Process and Content, Context, and Culture. *Educational Studies in Mathematics, 25*, 341-373.

Boaler, J. (1994). When Do Girls Prefer Football to Fashion? An Analysis of Female Underachievement in Relation to "Realistic" Mathematics Contexts. *British Educational Research Journal, 20*(5), 551-664.

Borasi, R. (1986). On the Nature of Problems. *Educational Studies in Mathematics, 17*, 125-141.

Brown S.B., Collins, A & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher 18* (1), 32-42.

Burton, L. (1993). Implications of constructivism for Achievement in Mathematics. In J.A Malone & P.C.S Taylor (Eds.). *Constructivist Interpretations of Teaching and Learning Mathematics* (pp.7-14). Perth Western Australia: National Key Centre for School Science and Mathematics.

Cambridge International Dictionary of English. (1996). Cambridge: Cambridge University Press.

Clarke, D.J. & Helme, S. (1998). Context as construction. In O. Bjorkqvist (Ed.), *Mathematics Teaching from a Constructivist Point of View.* Vasa, Finland: Faculty of Education, Abo Akademi University.

Cognition and Technology Group at Vanderbilt. (1990). Anchored Instruction and Its Relationship to Situated Cognition. *Educational Researcher, 19*(6), 2-10.

Cooper, B. (1992). Testing National Curriculum Mathematics: Some Critical Comments on the Treatment of 'Real' Contexts for Mathematics. *Curriculum Journal, 3*(3), 231-244.

Cooper, B. & Dunne, M. (2000*). Assessing Children's Mathematical Knowledge: Social Class, Sex, and Problem-solving*. Buckingham, Philadelphia: Open University Press.

Cumming, J.J. & Maxwell, G.S. (1999). Contextualising Authentic Assessment. *Assessment in Education, 6*(2), 177-194.
Fitzpatrick, R., & Morrison, E. J. (1971). Performance and Product Evaluation. In R. L. Thorndike (Ed.), *Educational Measurement* (2 ed., pp. 237-270). Washington, DC: American Council on Education.
Gerofsky, S. (1996). A Linguistic and Narrative View of Word Problems in Mathematics Education. *For the Learning of Mathematics, 16*(2), 36-45.
Greer, B. (1992). Multiplication and Division as Models of Situations. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 276-295).
Greer, B. (1993). The Modeling Perspective On Wor(l)d Problems. *Journal of Mathematical Behavior, 12*, 239-250.
Hambleton, R.K. & Murphy, E. (1992). A Psychometric Perspective on Authentic Measurement. *Applied Measurement in Education, 5*(1), 1-16.
Harmon, M., Smith, T.A., Martin, M.O., Kelly, D.L., Beaton, A.E., Mullis, I.V.S., Gonzalez, E.J. & Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA, USA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
Herman, J. (1997). Assessing New Assessments: How Do They Measure Up?. *Theory Into Practice, 36*(4), 196-204.
Hughes, S. (1993). What Is Alternative/Authentic Assessment and How Does It Impact Special Education?. *Educational Horizons, 72*(1), 28-35.
Illingworth, M. (1996). *Real-life Math Problem Solving*. NY: Scholastic Inc.
Kane, M., Crooks, T. & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.
Kirst, M. (1991). Interview on Assessment Issues with Lorrie Shepard. *Educational Researcher, 20*(2), 21-23, 27.
Lajoie, S.P. (1995). A Framework for Authentic Assessment in Mathematics. In T.A. Romberg, *Reform in School Mathematics and*

Authentic Assessment (pp.19-37). Albany, New York: State University of New York Press.

Lesh, R., & Lamon, S.J. (1992). Assessing authentic mathematical performance. In R. Lesh & S.J. Lamon (Eds.), *Assessments of Authentic Performance in School Mathematics* (pp. 17-62). Washington, DC: American Association for the Advancement of Science.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher, 20*(8), 15-21.

Maker, C.J. (1994). Authentic Assessment of Problem Solving and Giftedness in Secondary School Students. *The Journal of Secondary Gifted Education, 6*(1), 19-29.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.

National Council of Teachers of Mathematics. (1995*). Assessment Standards For School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Nesher, P. (1980). The Stereotyped Nature of School Word Problems. *For the Learning of Mathematics, 1*, 41-48.

Newmann, F.M. & Archbald, D.A. (1992). The Nature of Authentic Academic Achievement, in: H. Berlak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven & T.A. Romberg (Eds.), *Toward a New Science of Educational Testing and Assessment*: Albany, NY: State University of New York Press.

Newmann, F.M. (1997). Authentic Assessment in Social Studies: Standards and Examples. In G.D. Phye (Ed.), *Handbook of Classroom Assessment: Learning, Adjustment and Achievement*. San Diego, Ca: Academic Press.

Newmann, F.M. (1998). An Exchange of Views on “Semantics, Psychometrics, and Assessment Reform: A Close Look at ‘Authentic’ Assessments”. *Educational Researcher, 27* (6), 19-20.

Niss, M. (1992). Applications and Modeling in School Mathematics - Directions for Future Development. In I. Wirzup & R. Streit (Eds.), *Developments in School Mathematics Education Around the World,*

Vol 3. University of Chicago School Mathematics Project. Chicago: National Council of Teachers of Mathematics.
Office of Technology Assessment, U.S. Congress. (1992*). Testing in American Schools: Asking the Right Questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.
Organisation for Economic Co-operation and Development. (1999). *Measuring Student Knowledge and Skills. A New Framework for Assessment*. Paris: OECD.
Oxford Advanced Learners Dictionary of Current English. (1988). Oxford: Oxford University Press.
Palm, T. (2001). *Word Problems as Simulations of Real World Task Situations: A Proposed Framework.* Manuscript submitted for publication.
Reusser, K. (1988). Problem Solving Beyond the Logic of Things: Contextual Effects on Understanding and Solving Word Problems. *Instructional Science, 17*, 309-338.
Reusser, K. & Stebler, R. (1997). Every Word Problem Has a Solution – the Social Rationality of Mathematical Modeling in Schools. *Learning and Instruction, 7*(4), 309-328.
Romberg, T.A. (1995). Preface. In T.A. Romberg (Ed.), *Reform in School Mathematics and Authentic Assessment*. Albany, New York: State University of New York Press.
Ryans, D.G. & Frederiksen, N. (1951). Performance Tests of Educational Achievement. In E.F. Lindquist (Ed.), *Educational Measurement*. Washington, DC: American Council of Education.
Schack, G.D. (1994). Authentic Assessment Procedures for Secondary Students' Original Research. *The Journal of Secondary Gifted Education, 6*(1), 38-43).
Shepard, L.A. & Bliem, C.L. (1995). Parents' thinking about standardized tests and performance assessments. *Educational Researcher, 24*(8), 25-32.
Solano-Flores, G. & Shavelson, R.J. (1997). Development of Performance Assessments in Science: Conceptual, Practical, and Logistical Issues. *Educational Measurement: Issues and Practice, 16*(3), 16-25.

Sowder, L. (1989). Searching for Affect in the Solution of Story Problems in Mathematics. In D.B. McLeod & V.M. Adams (Eds.), *Affect and Mathematical Problem Solving: A New Perspective.* New York: Springer-Verlag.

Stenmark, J.K. (ed.), (1991). *Mathematics Assessment. Myths, Models, Good Questions, and Practical Suggestions*. Reston, VA: NCTM.

Swan, M. (1993). Improving the Design and Balance of Assessment. In M. Niss (Ed.), *Investigations into Assessment in Mathematics Education, An ICMI* Study (p.195-216). Dordrect, The Netherlands: Kluwer Academic publishers.

Terry, C.A. & Pantle, T.T. (1994). Authentic Assessment: Reducing the Fear and Trembling. *The Journal of Secondary Gifted Education, 6*(1), 44-51.

Terwilliger, J. (1997). Semantics, Psychometrics, and Assessment Reform: A Close Look at "Authentic" Assessments. *Educational Researcher, 26* (8), 24-27.

Terwilliger, J. (1998). Rejoinder: Response to Wiggins and Newmann. *Educational Researcher, 27* (6), 22-23.

Van den Heuvel-Panhuizen, M. 1996. *Assessment and Realistic Mathematics Education*. Freudenthal Institute, Utrecht: Technipress, Culemborg.

Verschaffel, L., De Corte, E. & Lasure, S. (1994). Realistic Considerations in Mathematical Modeling of School Arithmetic Word Problems. *Learning and Instruction, 4*(4), 273-294.

Verschaffel, L., & De Corte, E. (1997). Word Problems: A Vehicle for Promoting Authentic Mathematical Understanding and Problem Solving in the Primary School?. In T. Nunes & P. Bryant (Eds.), *Learning and Teaching Mathematics: An International Perspective* (pp. 69-97). Hove: Psychology Press.

Wiggins, G. (1989). A True Test: Toward More Authentic and Equitable Assessment. *Phi Delta Kappan, 70*, 703-713.

Wiggins, G. (1993). Assessment: Authenticity, Context, and Validity. *Phi Delta Kappan, 75(3)*, 200-214.