Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 13, Number 4, April 2008 ISSN 1531-7714

Performance Assessment and Authentic Assessment: A Conceptual Analysis of the Literature

Torulf Palm, *Umeå University, Sweden*

Performance assessment and authentic assessment are recurrent terms in the literature on education and educational research. They have both been given a number of different meanings and unclear definitions and are in some publications not defined at all. Such uncertainty of meaning causes difficulties in interpretation and communication and can cause clouded or misleading research conclusions. 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.

There are a number of ill-defined concepts and terms used in the literature on education and educational research. This is a problem for many reasons, and one of them is the difficulty of interpreting research results. There are several examples in the literature of loosely defined constructs that have been used differently in different studies, which have caused different results and in turn clouded and caused misleading conclusions (see e.g. Schoenfeld, 2007; Wiliam, 2007). The diversity of meanings also makes communication and efficient library searches more difficult. Performance assessment and authentic assessment are two concepts that have been given a multitude of different meanings in the literature and are used with different meanings by different researchers. In addition, they are sometimes only vaguely defined and sometimes used without being defined at all. This multitude of different meanings, especially in the light of the lack of clear definitions in some publications, makes it difficult for teachers and newcomers in the assessment research field to get acquainted with the research in this area. But it also causes misunderstandings and communicational problems among experienced researchers, which is evident from a debate in the Educational Researcher (Brandt, 1998; Newmann, 1998; Terwilliger, 1997, 1998; Wiggins, 1998). Furthermore, due to different histories of assessment practices the difficulties caused by the confusion about the meanings of these concepts may arise even more easily in situations involving international participation (such as actions taken based on readings of international research journals). The introduction of the term authentic assessment and the increase in use of the term performance assessment in theoretical school subjects seem to have come as a response to the extensive use of multiple-choice testing in the US. But since many countries do not have, nor have had, such an extensive use of multiple-choice testing many non-US researchers and practitioners do not share the experiences that led to these different meanings, which causes very different bases for interpreting the situation with all of the different (and sometimes vague) meanings. Indeed, a corresponding concept to performance assessment does not even exist in many countries.

The aim of this article is neither to present additional definitions nor to make judgments on existing ones. The intention with the article is to analyze the meanings given to the two concepts performance assessment and authentic assessment in the literature in an attempt to clarify the diversity as well as the similarities of the existing meanings. Such a survey may be helpful for communication about important assessment issues and also for further efforts of coming up with definitions that can be agreed upon, which for reasons mentioned above indeed would be desirable.

For these aims, it is important to acquire a full picture of the variety of meanings these concepts possess.

Most definitions of performance assessment seem to be subject-independent and therefore the section about this concept mostly deals with definitions not specific to a particular subject. Since performance assessment sometimes is described by its typical characteristics and sometimes by a more clear definition the section about performance assessment includes one subsection describing the characteristics that have been argued in the literature to be typical of performance assessments, and a subsequent subsection describing the different definitions. The latter subsection begins with an overview of different types of definitions that have been put forth and concludes with examples of definitions to exemplify the similarities and differences of the meanings of the definitions. Authentic assessment is treated in the following section. Definitions of authentic assessment are also often subject-independent, but not to the same extent as performance assessment. Therefore, both subject-independent and subject-specific definitions will be included. The subject mathematics will be used to exemplify the subject-specific definitions. The first subsection on authentic assessment provides a classification of different meanings, and is followed by two subsections with examples of definitions intended to clarify the classification.

Brief history

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, Crooks & Cohen, 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 consensus on the meaning of performance assessment has not been reached.

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). However, 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.

METHOD

In the search for definitions and descriptions of the concepts the ERIC database and the mathematical education database MATHDI from Zentralblatt für Didaktik der Mathematik were used. Searches were made for the terms "performance assessment", "authentic assessment", "authenticity" and "authentic" in the titles or in the abstracts of the publications included in the databases. The search was mostly restricted to publications written in English. The abstracts were scanned for indications that the publications did include some kind of definition of one or both of the terms. These publications were collected and the definitions were analyzed. In addition, the references in the collected publications were used to find other publications that included descriptions of the concepts of interest. The search for publications was terminated when abstracts and references most likely to include clear definitions had been analyzed and no new meanings seemed to appear in the additional publications collected. There is no feasible way of finding every definition of the concepts in the literature, and no such claims are made here. However, an extensive search has been made, and since in the end of the search no new meanings were detected as new references were collected, it is likely that most of the frequent meanings presented in the English written literature could be described by the developed categories.

The actual development of the taxonomy, that is, the choice and description of different categories of meanings, can be made in different ways, and especially categorizations made on different grounds may end up in slightly different taxonomies. For example, the analysis by Cumming & Maxwell (1999) of various ways in which authentic assessment is interpreted offers a different categorization than the categorization of meanings of authentic assessment provided in this paper. Their analysis was made on the basis of the learning theories underlying the different meanings of the concept. That is, it was based on the different interpretations of knowledge and learning that seemingly has led to variations in the constructions of authenticity and the implementation of authentic assessment.

The purpose of the categorization in this paper was to develop a description of the meanings attached to the concepts of interest that would reveal the features of the meanings as clear as possible. The meanings found in the collected publications were analyzed to find categories that would describe the features of these meanings in such a way that the similarities and differences between different meanings would appear distinctly. Examples of definitions to exemplify different set of meanings were chosen on the basis of their possibilities to reveal the characteristics of the specific sets of meanings and the differences to other sets of meanings.

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.

Characteristics

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 judgments 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.

Categories of definitions

The definitions of performance assessment put forth are of a different kind than the characteristics. 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 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, the requirements by the Office of Technology Assessment, U.S. Congress (OTA, 1992) 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. Many assessments that would be regarded as performance assessment by the definition of the OTA would not be considered to be performance assessment with the requirements of Kane et al. 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. Since some of the simulation-centered definitions require special equipment use, it is also clear that there are 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 restraints in the examinee's access of tools, collaboration, and literature and so forth than the restraints in the simulated real situation. It can also be interpreted as an assessment administered for classroom use, demanding only, for example, traditional mathematics word problems requiring short student-constructed responses.

Examples of definitions

In the following a guided tour over different definitions is undertaken to exemplify the similarities and differences between the definitions categorized in the two main categories of definitions mentioned above. In the definition made by the Office of Technology Assessment, U.S. Congress, (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*" (p. 30).

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 et al. (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; "*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*" (Kane et al., 1999, p. 6-7). Thus, that an assessment involves performances that are valid indicators of the performances of interest does not suffice to be considered a performance assessment.

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.

Mostly when performance assessment is discussed generally or for specific subjects a subject-independent definition is called upon. However, it does exist subject-dependent definitions. For example, Solano-Flores & Shavelson (1997) relates the performance of interest to what scientists do when they define science performance assessment 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).

AUTHENTIC ASSESSMENT

As in the case with performance assessment authentic assessment can mean almost anything. The first subsection includes a description of perspectives and foci taken 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 the next two subsections the different perspectives and foci are exemplified through a number of definitions of authentic assessment. The former of these two subsections deals with general definitions and the latter subsection deals with definitions in the special case of school mathematics. The ambition has been to select illustrative examples of the types of meanings that pertain to the identified 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 the term authentic assessment. However, since tasks are the building blocks and 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 assessment tasks are considered as well.

Classification of meanings

In the Cambridge advanced learner's Dictionary (online) something that is authentic is explained as "*it is real, true, or what people say it is*". In relation to assessment, the explanation in the dictionary 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 issues 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 issue:

1. **Life beyond school.** With this perspective similarity to life beyond school is emphasized. This can include the requirement that students during the assessment 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.

2. **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.

3. **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 issue "*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 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 (Focus 1 combined with Perspective 1), meet curricula goals (Focus 1 combined with Perspective 2), or are effective in the learning process (Focus 1 combined with Perspective 3).

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 (Focus 2, Perspective 1) or in ordinary classroom practice (Focus 2, Perspective 2). The third perspective, learning and instruction, would, combined with this focus on 'conditions', require 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 the particular school subject, for example mathematics, in which the task is given. 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 focus is always combined with Perspective 1, but sometimes accepts other school subjects than mathematics to also be included in this perspective).

The above 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. But it does mean that these perspectives and foci represent different frames of reference chosen in defining authentic assessment, resulting in different meanings of the concept.

Examples of definitions

The concept of authentic assessment is a much more recent term than performance assessment. According to Cumming & Maxwell (1999, p. 178) and a discussion in Educational Researcher (e.g. Wiggins, 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)

In authentic assessment the mastery defined by the concept of authenticity is assessed. This means that in authentic assessment students should construct knowledge. 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.

Thus, the defining features of authentic assessment are the specific cognitive processes (disciplined inquiry) and products (knowledge beyond the mere reproduction of presented knowledge) considered important in the perspective of life beyond school. But in addition, Newmann & Archbald (1992) also argue that the students' working conditions and other assessment characteristics are important for the possibilities of eliciting these processes and products. They specify a number of such conditions, also related to the perspective of Life beyond school, which include that the students have the opportunities to collaborate and that the assessment has criterion-based standards. 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.

Wiggins' perspective of authentic assessment is also 'life beyond school' and in addition to 'processes and products' he also emphasizes 'conditions'. He does not specify the 'processes and products' in the same way that Archbald and Newmann do but 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. Organisation for Economic Co-operation and Development (1999) below).

Shepard (as quoted by Kirst, 1991) has learning and curriculum as perspective in her approach to authentic assessment, which changes the meaning of the concept. 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 performances, 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.

But also with the same perspective the meaning of authentic assessment may differ. Messick (1994) takes the curriculum perspective in a broad meaning. 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). 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 (Embretson, 1983; Messick, 1989). That is, evidence should be sought that the presumed sources of task complexity are indeed reflected in task performance and that the complex skill is captured in the test scores with minimal construct underrepresentation.* (Messick, 1994, p. 20)

Shifting the main perspective to learning and instruction significantly changes the meaning of authentic assessment. According to Schack (1994) authentic assessments include that the assessments "*give students both feedback upon completion" as well as "guide their work along the way*" (p. 39).

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' suggests a focus on the assessment developer's intention with the tasks or assessments. As a consequence the tasks would not necessarily have to be experienced as valuable by the students as long as this was the test developer's intention. In contrast, a definition requiring that the students really do experience the tasks as valuable put much harder demands on the assessment development. The difference may at first glance be seen as a trifling technicality but may prove to be crucial in developing, evaluating and revising assessments as well as for the meaning of the concept of authentic assessment.

Examples of definitions specific to school mathematics

In defining authentic assessment in the special case of school mathematics some authors call upon a general definition. Other authors include mathematics-specific meanings in a definition. Stenmark (1991) is an example of the latter. She specifies the influence of the specific discipline of mathematics on the definition of authentic assessment tasks in mathematics. Focusing on 'Processes and products' and taking the perspective of Life beyond school 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)

Another example of the influence of the specific nature of mathematics is 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 (taking the perspective of 'Learning and Instruction'). These principles involve the requirement of alignment with the *Curriculum and Evaluation Standards* by the National Council of Teachers of Mathematics (NCTM, 1989), which constitutes the mathematics specific standards to which the students' learning are to be directed. 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 definition in the mathematical literacy framework of the OECD's Programme for Student Assessment, PISA, (Organisation for Economic Co-operation and Development, 1999) is an example of a definition in which the focus is on the figurative context and not on specific 'processes and products' nor on students' working conditions. The issue is that the figurative context truthfully describes a situation from real life that has occurred or might happen. A task seems to be regarded as authentic if its figurative context, the situation described in the task, is authentic, and this context is authentic if "*it resides in the actual experiences and practices of the participants in a real-world setting*" (p. 51).

CONCLUSIONS AND DISCUSSION

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 from the 1990s and onwards (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.

From the analysis of different definitions of performance assessment it is clear that some of the definitions share important properties. At the same time it is also evident that performance assessment can mean almost anything. 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 the students' work include non-written performance. Furthermore, there is a gap between the descriptions of performance assessment that are characterized by the assessment's characteristics, in the sense of typical properties, and the descriptions of performance assessment that are characterized by a definition. 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. It is obvious that student-constructed response (beyond selecting from a set of ready-made answers) is a prerequisite for students' extended communication, and 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.

Authentic assessment is often associated with assessment emulating real life task situations, but also possesses meanings such as assessment aligned with curriculum and assessment that effectively supports learning. The 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 authentic assessment. 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 (Schack, 1994). In addition, the descriptions of authentic assessment are often quite indistinct and sometimes even contradictory within the same publication.

In a comparison of the meanings given to performance assessment and authentic assessment the analysis shows that they share some of the meanings given to them. Several of the 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 the particular school subject. 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. Different purposes of reform and different views on for example knowledge, learning and assessment have probably contributed to the diversity of meanings. The choice of term (authentic) may also have added to the difficulties of maintaining a reasonably well-defined meaning of this concept. 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).

It is of course always important to provide clear definitions of terms used in research presentations. However, in the light of the extreme variation of views on the concepts discussed in this paper, and the difference in assessment practice (and history of assessment practice) around the world it seems that such a clarification is particularly important for these concepts and even more so when the publication is aimed at an international audience. 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. The description in this paper of both similarities and differences of the meanings of these concepts may be useful in communication involving these concepts, both from the writer's and from the reader's perspective.

REFERENCES

Airasian, P.W. (1994). *Classroom assessment (2nd ed.).* 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). Baltimore, Maryland, John Hopkins University Press.

Brandt, R. (1998). Research news and comment: An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments". *Educational Researcher, 27* (6), 20.

Cambridge advanced learner's dictionary. The word was looked up January 4, 2008, at http://dictionary.cambridge.org

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.

Cumming, J.J. & Maxwell, G.S. (1999). Contextualising authentic assessment. *Assessment in Education, 6*(2), 177-194.

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.

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.

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. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

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). Research news and comment: An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments". *Educational Researcher, 27* (6), 19-20.

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.

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.

Schoenfeld, A.H. (2007). Method. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 69-107). Charlotte, NC: Information Age Publishing.

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.

Stenmark, J.K. (ed.), (1991). *Mathematics assessment: Myths, models, good questions, and practical suggestions.* Reston, VA: NCTM.

Terwilliger, J. (1997). Semantics, psychometrics, and assessment reform: A close look at "authentic" assessments. *Educational Researcher, 26* (8), 24-27.

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

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.

Wiggins, G. (1998). Research news and comment: An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments". *Educational Researcher, 27* (6), 20-22.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1051-1089). Charlotte, NC: Information Age Publishing.

Note

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

Citation

Palm, Torulf (2008). Performance Assessment and Authentic Assessment: A Conceptual Analysis of the Literature. *Practical Assessment Research & Evaluation*, 13(4). Available online: http://pareonline.net/getvn.asp?v=13&n=4

Author

Torulf Palm
Department of Mathematics, Technology and Science Education
Umeå University
901 87 Umeå
Sweden

Email: torulf.palm [at] math.umu.se