Olfos, R., & Zulantay, H. (2007). Reliability and Validity of Authentic Assessment in a Web Based Course. *Educational Technology & Society*, 10 (4), 156-173.

Reliability and Validity of Authentic Assessment in a Web Based Course

Raimundo Olfos
Pontificia Universidad Católica de Valparaíso. Mathematic Institute, Valparaíso, Chile// Raimundo.olfos@userena.cl

Hildaura Zulantay
Universidad de La Serena. Mathematic Departemant, La Serena, Chile // hzulantay@yahoo.com

ABSTRACT

Web-based courses are promising in that they are effective and have the possibility of their instructional design being improved over time. However, the assessments of said courses are criticized in terms of their validity. This paper is an exploratory case study regarding the validity of the assessment system used in a semi presential web-based course. The course uses an authentic assessment system that includes online forums, online tests, self-evaluations, and the assessment of e-learner processes and products by a tutor, peers, and an expert. The validity of the system was checked using internal and external criteria. The results show that the authentic assessment system addresses technical problems, especially regarding reliability of instruments. Some suggestions are proposed to strengthen authentic assessment in web-based courses and to test it.

Keywords
Authentic assessment, Reliability and validity, Web based instruction, Evaluation

Introduction

Web-based instruction offers flexibility, and it is becoming increasingly popular (Walles, 2002). Some experiences show Web-based instruction is as effective as classroom instruction (White, 1999; Ryan, 2001; Tucker, 2000). Moreover, Wilkerson and Elkins, (2000) concluded that students felt that web-based instruction was as effective as that in traditional courses. The use of the Web for instruction is in an early stage of development, and its effectiveness may not yet be fully known (Downs et al., 1999). In an extensive study including 47 assessments of web-based courses, Olson and Wisher (2002) assert that web-based instruction appears to be an improvement over conventional classroom instruction.

To Phipps and Merisotis (1999) quality of the research in distance learning courses is questionable: much of the research does not control for extraneous variables; the validity and reliability of the instruments used to measure student outcomes and attitudes are questionable; and many studies do not adequately control for the feelings and attitudes of students. Attitude is a learned predisposition to consistently respond to a given social object (Segarra et al. 1997); for instance, technology tools.

Authentic Assessment

The use of new technology raises issues related to pedagogy, content, and interaction. As these issues are addressed, there needs to be a subsequent alteration in the type of assessment used on such courses and the associated procedures (Walles, 2002).

A new approach to evaluation is authentic assessment. This modality connects teaching to realistic and complex situations and contexts. Also called performance assessment, appropriate assessment, alternative assessment, or direct assessment; authentic assessment includes a variety of techniques such as written products, portfolios, check lists, teacher observations, and group projects.

According to Herrington and Herrington (1998) authentic assessment occurs within the context of an authentic activity with complex challenges, and centers on an active learner that produces refined results or products, and is associated with multiple learning indicators. It includes the development of tests and projects (Condemarín, 2000).

The AAS not only evaluates the products, but also the processes involved. It is a process that monitors the learner's progress using a variety of methods, such as observation records, interviews, and evidence gathering. It is consistent with Vygotsky's dynamic assessment in that it is a mediated process in which social interaction stimulates the learning process. It also allows for collaborative learning.

Authentic assessment of educational achievement directly measures actual performance in the subject area. It was developed as a result of criticism of multiple-choice tests, which usually only provide a superficial idea of what a student has learned and do not indicate what a student can do with what was acquired (Aiken, 1996). Authentic assessment can provide genuine accountability. All forms of authentic assessment can be summarized numerically, or put on a scale, to make it possible to combine individual results and to meet state and federal requirements for comparable quantitative data (National Center for Fair and Open Testing, 1992). One method to this end, according to Wiggins (1998), is the use of rubrics, which are sets of criteria that evaluate performance. Points are assigned according to how well each criterion is fulfilled, and are then used to provide the quantitative values.

The validity of Authentic Assessment

Aiken (1996) plainly indicates the difficulty in establishing the validity and reliability of any authentic assessment. Schurr (1999) states that a disadvantage of authentic assessment is the difficulty, maybe even the impossibility, of obtaining consistency, objectivity, and/or standardization in its results. Wolf (cited by Herrington and Herrington, 1998), sees the problem as a tradeoff between validity and reliability. In the same vein, Linn, Baker and Dunbar (1991), hold that performance-based assessments are valid in terms of consequence, impartiality, transference, content coverage, cognitive complexity, significance, judgment, cost, and efficiency; consequently, reliability, understood as the stability of the results, is difficult to obtain.

Linn et al. state that "a greater dependence of critical judgments on the performance of tasks" is inevitable. This is a problem for large-scale assessments, but not for the smaller and more specific contexts to be found in superior education (Gipps, 1995), or in those cases in which the student is evaluated through class activities, where learning and evaluation are essentially one and the same (Reeves and Okey, 1996).

There are also problems with reliability in standardized tests. Gipps (1995) mentions national tests in the UK with reliability rates in posttests that are lower than those obtained in performance tests. Reeves and Okay (196) state that authentic assessment takes place within a real-world context, where generalizations are of little value and, therefore, reproducibility should not be a concern. Young (1995) adds that assessment needs can be perceived in a more functional way, and assessments can be validated in terms of their value in the real world rather than for their stability as instruments (Herrington and Herrington, 1998).

Authentic assessment in web based courses

Computing-mediated distance education introduces extraneous factors that could affect the validity of the course assessment system. One of these factors is identified as the usability of the web site. Usability deals with how well a system satisfies user needs and requirements. It applies to all the aspects of a system with which a user might interact, including installation and maintenance procedures. It is usually associated with aspects such as ease of learning, user friendliness, ease of memorizing, minimal errors, and user satisfaction. (Sánchez, 1999, 2000). Another factor is attitude. According to Ceballos et al. (1998), only when there is a favorable attitude toward the TICs, can an e-learner effectively face learning tasks; therefore a web-based assessment system requires a positive attitude from the users to show their full potential. In other words, the degree of effectiveness of the assessment system could be affected by a negative attitude towards the TICS by some of the e-learners.

Literature provides scarce information about authentic assessment in web-based courses. It only refers to case studies and similarities. Weller (2002) examines technical barriers in the assessment process of a web-based course, and points out the tension between individuality and robustness in submissions and the detection of plagiarism. Clarke et al. (2004) state that feedback to students on their assignments is an essential activity. They point out that tutorial support must be a crucial part of good distance course, where emails should be considered as a non-intrusive means of communication. Orde (2001) offers some suggestions to develop online courses, for instance: to consider a description of learners; to provide readily available technical support; to eliminate group activities easily done face-to-face; and to record and grade interactions such as e-mail and group discussion contributions. To Orde, the testing portion of CourseInfo requires that each quiz item be individually entered and submitted. If this feature of the software is used, the ID students advise allowing multiple attempts. To Orde, formative evaluation is an essential component and necessary to online course development.

Gatlin and Jacob (2002) discuss advantages of digital portfolios as part of one university's authentic pre-service teacher assessment. Chang (2002) asserts that web based learning portfolios are useful for students to obtain feedback from other students. Collins et al (2001) state that the design of assessment for web based courses should measure student participation, and the development of skills.

Literature recommends centering authentic assessment on individual tasks, but also connecting it to real life, including interactive work among peers. Fenwick and Parsons (1997) assert that effective assessment must be intricately woven throughout the teaching-learning process. Collaborative learning activities enable subjects to share their abilities and limitations, providing a better quality product than one that is the mere sum of individual contributions. Group interactions, following the indications for collaborative work given in a course, facilitate vicarious learning, which is hard for some subjects to experience if they do not interact with their peers (Bandura and Walters, 1963; Vygotsky, 1985).

The purpose of this study

The purpose of this study is to analyze the validity of the Authentic Assessment System (AAS) used in a web-based online course. Leading questions of this study are: Is there consistency between AAS results and effectiveness of products in a real life context? How much do AAS results correlate with external individual learning criteria? What evidence can be gathered about the reliability and validity of AAS's different phases? "Do external factors such as web usability and participants' attitude invalidate the AAS design?

Method

Subjects

Twenty-eight teachers participated, 13 primary school teachers and 15 secondary school math teachers, who taught different levels of math, from 5th to 10th grade. All except one of the participants worked in schools in two nearby cities or their surroundings. The cities had a population of approximately 120,000 inhabitants each. None of the students had previous experience with distance learning.

Setting

A.- The course. This study took place within the context of the twelve-week course called "Didactics of Elementary Algebra" with 89% effectiveness, offered to mathematics teachers. Twenty-five of the 28 teachers –hereinafter called students- passed the course and met AAS requirements. It was a semi presential course. The students did most of the activities by means of distance learning. To do so, the students had access to a web site with the course contents and computer resources such as e-mail, forums, FTP, and online tests in order to interact with the rest of the students and with the tutor in order to meet the course's evaluation requirements. The course included three classroom activities and an authentic assessment system made up of eight internal examinations.

The course was managed by a coordinator and led by a tutor, both of whom were academicians with experience in distance learning. The students used computers as a means of communication with their peers as well as with the tutor. In addition, computers were used to do the course assignments as well as facilitating the search for and creation of didactic material.

The course objectives were: a) to develop the capacity to design, implement, and evaluate didactic units regarding elementary algebra, b) strengthen the capacity to work collaboratively, design and evaluate the use of didactic units centered on elementary algebra, and c) develop Internet skills with the aim of strengthening professional skills and efficiency in teaching elementary algebra.

The first unit, "Internet tools" (1 week), included a classroom session in which surfing the Internet, using e-mail, and the website's communications services were reviewed. The second unit, "Didactic Design Concept" (2 weeks), showed the students how to characterize the unit's terms, didactic design, lessons and activities, by establishing a basic language and conceptual framework upon which to develop the didactic designs. An individual evaluation was included. It was called "Design forum" (I-1) referring to units 1 and 2. The third unit, "Professional Knowledge" (2 weeks), introduced specific concepts on types of learning, curricular conceptions, and learning evaluation. The unit included an individual examination called "On-line test" (I-2) referring to course units 1, 2, and 3.

The fourth unit, "Creating a Didactic Design" (4 weeks), guided the student in creating a didactic design. In this unit, collaborative work in small groups was realized, in which decisions were negotiated and roles were assigned to the participants with the aim of achieving common objectives. Three evaluations were done: assignment of “individual contributions" (I-3), the group creation of a "preliminary didactic design" (I-4), and the creation of a "didactic design" (I-5). During this period, the fifth week, a classroom session took place to get feedback from the collaborative work. Material was shared and the work was checked ahead of time. The fifth unit "Validation of the Designs" (2 weeks) provided the guidelines for testing the didactic designs with schoolchildren. In addition, it provided the guidelines for creating a final report that included a commented version of the didactic design subsequent to its application. This unit included three evaluations: the "final report" (I-6), a self-evaluation of the design through the "Own Design Forum" (I-7) and self-evaluation or "personal evaluation guideline" (I-8). The course ended after 12 weeks with the third classroom session. In this session, the designs were shared.

B.- Authentic Assessment System. The assessment system regarding the students' learning and group production included the eight examinations mentioned (I-1 to I-8), which were structured following the four criteria summarized by Herrington and Herrington (1998).

- The authentic assessment took place in a real context. It was connected with the students' daily professional performance (Meyer, 1992; Wiggins, 1993; Reeves & Okey, 1996).
- In the authentic assessment, the students assumed a leading role in collaboration with their peers (Linn et al., 1991; Kroll, Masinglia y Mau, 1992) displaying their performance and presenting their products (Wiggins, 1989, 1990, 1993).
- In the authentic assessment, the evaluation activities were authentic. In other words, they were inextricably integrated with course learning activities and the student's daily professional activity (Young, 1995; Reeves & Okey, 1996). The evaluation activities corresponded to unstructured, complex challenges which demand students' own opinions (Wiggins, 1993; Linn et al., 1991; Torrance, 1995).
- In the authentic assessment, multiple indicators (Lajoie, 1991; Linn et al., 1991) and several criteria were used to grade the variety of requested products (Wiggins, 1993; Lajoie, 1991; Resnick & Resnick, 1992).

Table 1. Summarizes structure and content of each instrument used in an AAS

Procedure	Description
I-1 Forum Design	Participation in the forum, Dichotomist (does/ does not)
I-2 On Line Test	30 graded items attending 6 content dimensions, 3 levels of complexity.(Resent options were admitted)
I-3 Individual Contributions	2 records referred to e-mails sent Dichotomist:Sent pertinent messages or not
I-4 Preliminary Design	5 graded records referred to pre design attributes. Rubric based
I-5 Didactic Design	11 graded items referred to Design. Rubric based
I-6 Final Report	9 graded items referring to the design application. Rubric based
I-7 Own Design Forum	1 numeric record summarizing didactic design quality opinion
I-8 Personal Evaluation	20 Likert scale items. Knowledge, abilities and feelings acquired

Individuals were basically tested on knowledge domain. The first part of the course provided instrumental knowledge about how to build didactic design, how to work in a collaborative approach, and how to use e-mails and the web site. Individuals were required to answer various tests and were also asked to report their learning perception during the process. The second part of the course focused on a productive approach. Participants formed small groups to elaborate didactic designs. Accordingly, evaluation procedures were applied to partial products of the didactic design building.

Table 2. Assignment evaluation responsibilities and the weight of each procedure in relation to the assessment system

Responsible of assignment	Evaluated	Weight
I-1 Automatic assignment	Individual	5
I-2 Automatic assignment.	Individual	5
I-3 Automatic assignment. Pertinence was decided by tutor	Individual	5
I-4 Tutor does assignments.	Group Group	10
I-5 Peers: two or more colleges do assignments	Group Group	25
I-6 Expert does assignments	Individual	25
I-7 Assignment by participants (self evaluation)		5
I-8 Self evaluation learning		10

Research Design

This single case study focused on multiple sources as Yin (1984), Stake (1995) and Weiss (1997) have recommended. Tests, archival records, and participants' perspective were recorded. According to Yin (1994), this case study design was used to describe the AAS implemented in a specific web based course, and to analyze several variables related to a small number of subjects.

The study considered individuals and small groups as units of analysis, according to an embedded case study design (Yin, 1994). Data were connected to indicators by means of correlations and triangulation analysis. Traditional alpha 0.5 and 0.01 were used.

Both the evidence of validity based on criteria of judges, parallel instruments, and non obstructive data, and the evidence of reliability obtained from indices of association between equivalent items, and reiterated measurements, are sustained in correlations. In effect, the Rho of Spearman coefficient for nonparametric data, index r of Pearson for variables of normal distribution, and the value alpha of Cronbach for groups of variables, express a correlation or degree of association between the measured variables. These univariated measurements are used in exploratory studies because they are simple and they aid in elaborating integrated models. 5% and 1% of the critical values considered in this study are used habitually in education with cuasi-experimental designs, unlike the values such as 0.1%, or lower, used in experimental designs or in the area of medicine.

Instruments

This study used records of the eight AAS procedures, and three additional data sources to validate the AAS.

The eight AAS instruments

AAS was made up of 8 procedures associated with their respective instruments. When using the Cronbach coefficient as an indicator of AAS's reliability, we got a positive, but not significant value ($\alpha(26)=.22$, $p>.05$), which was expected because AAS measurements did not refer to a one-dimensional variable.

Table 3. Correlation of each instrument with the rest of the AAS

Internal Instruments	Correlation[1] (one-tailed)
I-1 Forum Design contribution	*rho*(26) = 0,37 *p*<.05
I-2 On Line Test (resent options)	*rho*(26) = 0.29 *p*>.05
I-3 Individual Contributions mailed	*rho*(28) = 0,50 *p*<.001
I-4 Preliminary Design check list	*rho(28)* = 0,61 *p*<.001
I-5 Didactic Design check list	*rho*(26) = -.02 *p*>.05
I-6 Final Report check list	*rho*(26) = 0.11 *p*>.05
I-7Own Design Forum opinion	*rho*(26) = 0.47 *p*<.001
I-8 Self Evaluation Guideline	*r*(28) = 0,24 *p*<.05

The eight AAS procedures measured different types of individual learning and group work. The eight correlations obtained between each instrument and the rest of the assessment system, as internal consistence indicators, were low, making it evident that the measurements were heterogeneous. See table 3.

The online test (I-2) reached Cronbach $\alpha(26)=.88$, which was used as a reliability measure. The Likert scale (I-8) reached Cronbach $\alpha(13)=.96$, $p<.001$ which represented internal consistence between the items.

The three sources of external data to AAS. The external data sources were non-obstructive records, posttest results, and answers to questionnaires. Data were collected before, during, and after the course.

E-1. Non-obstructive records. On the one hand, the Final Report (I-8) provided data of design effectiveness with schoolchildren in different classrooms. An overall index of design effectiveness was calculated by means of schoolchildren performance after Design application (E-1 a). On the other hand, the Web site provided facilities to get data about students' participation in the course. E-mails sent to the tutor were counted (E-1 b).

E-2. Posttest. The test included 16 multiple-choice items: 6 didactic design items, 4 ITech items, and 6 collaboration items (E-2), to cover the main contents of the course.

E-3. The questionnaires. There were three questionnaires. One referred to learning and motivation feelings throughout the course (E-3 a), another to the students' attitude towards the AAS (E-3 b), and the last, to the usability of AAS aspects of the website (E-3 c). These data were collected during the second indoor session.

Table 4 shows the external instruments. The table includes indexes of posttest (E-2), questionnaires (E-3 a), (E-3 b) and (E-3 c).

Table 4. External procedures

Evaluated		Items Description	Source	Reliability
E-1a	Group	Single record: design usage effectiveness	Student Final Report	---
E-1b	Individual	Single record: e-tools used during the course	Data of Web site use	----
E-2	Individual	10 (4 didactic design, 4 ITech , 6 collaboration)	Multiple choice test	rho(25) =.67
E-3a	Individual	10 multidimensional opinions (learning, motivation)	Likert scale	----
E-3b	Individual	24 items: students' attitude to the AAS	Likert scale	α=.62 (n=27)
E-3c	Individual	23 items: evaluation aspects of website usability	Likert scale	α=.74 (n=22)

External Instruments Description

E-1a. Designs effectiveness: Data about design effectiveness were included in the final students' report. Course students collected information referring to schoolchildren's learning through the Design application. This information was computed by means of schoolchildren's grades obtained during design application in their classes.

E-1b. Posttest: A pre-test was created according to the specification table regarding the course's three main concentrations of contents. Only items with a difficulty of more than 50% were assigned to the post-test. Items similar to the 8 items selected from the pre-test were designed, and in this way the posttest was made up of 16 similar items. The reliability of the post-test, calculated by the method of dividing into halves, yielded rho(25)=0.67.

E-3b. Likert scale about attitude: According to Anastasi (1982), attitude, in terms of the tendency to react favorably or unfavorably toward a certain class of stimuli, was determined by visible, both verbal and non-verbal, behavior. An attitude test toward the mathematics from The School Science and Mathematics Association Inc was adapted (Aiken, 1996). The items measured the student's attitude toward the course's authentic assessment system. The proposed model and original test, which established a balance between positive and negative attitudes, were applied. Twenty-four items were created, assigning a score on the Likert scale of 1 to 5 points, 1 being the lowest and 5 being the highest. Educators with experience in distance learning validated the content and classified the statements according to type, positive, or negative attitudes. The instrument's internal consistency was measured, thereby obtaining Cronbach α = .6 2 (n=27).

E-3c. Likert scale on usability: The scale was implemented as an indicator of the incidence of the usability of the web site in the course results. An adaptation of the Sanchez (2000) usability test was carried out, focusing on the measurement of the usability of the Website in relation to the course's authentic assessment system. The author used five proposed dimensions: learning, satisfaction, error, efficiency and memory factors. Twenty-three items were created. They were assigned a score on a Likert scale, of 1 to 5 points, one being the lowest grade and five being the highest. The test was subjected to the remote criticism of educational research specialists for the validation of its content. The instrument's internal consistency was measured with Cronbach's alpha, obtaining respectively for each one of the factors, the coefficients 0.68, 0.77, 0.84, 0.54, 0.89, and for the complete instrument, $\alpha = 0.74$ (n=22).

Procedures

The first two procedures

They refer to global indicators of AAS' validation: the relation with real context consequences, and with individual learning.

Consistency between AAS' results, and product effectiveness in real context. After designs were elaborated, materials for schoolchildren were reproduced to use in one or more classes. Then lessons were implemented and schoolchildren were evaluated. The averages of these evaluations were compared with the students' AAS results by means of Spearman's range correlation coefficient.

Measurement of degree of association between AAS results and individual learning criteria. A post-test was considered as an indicator of the students' learning. The posttest (E-2) was applied at the end of the course without prior warning; the students had not prepared for the test, and the results did not have any effect on the class grades. For this concurrent validity analysis, the students' AAS average was correlated with the post-test average.

What evidence can be gathered about AAS validity in either individual or group phases?

Analysis of the validity of AAS's Individual Component

The individual phase focused on instrumental learning. Three criteria were considered:.

The student's capacity to recognize the components of a didactic design. The student's capacity to recognize the components of a didactic design was measured in one of AAS's instruments as well as in the external post-test. This analysis compared the percentages of correct answers from the items referring to this capacity included on the posttest (E-2) as well as on the online test (I-2).

The student's capacity to surf the web site and send e-mails. This analysis about the students' capacity to navigate the web site and use e-mails was based on information gathered from the evaluations applied to the students and from non-obstructive information from the website. Specifically, two indicators of the students' capacity to surf the website and use e-mails were created. The first indicator was the sum of four AAS data, namely: a publication in "forum 1" (I-1), answer to an item in the "on-line test" (I-2), answers to two items in the "self-evaluation" (I-8). The second indicator was the sum of the data obtained from the instruments that were external to AAS, namely: the number of e-mails sent, with a maximum of three (E-1 b), an item from the "external questionnaire" (E-3 a), three items referring to the website usability (E-3 c), and three items from the post-test" (E-2). To create these indicators, each variable was adjusted to a scale of 0 to 1, without considering the values assigned to the AAS instruments. The AAS information was correlated with the indicator associated with external data.

Students' opinion on learning about didactic designs and IT. During the course, there were two opinion instruments that referred to the learning level attained during the course; One was part of the AAS (I-8) and the other was not (E-3 a). Although the instruments were administered at different times, one in the middle and the other at the end of the course, the degree of association between the answers of groups of equivalent items was used as an indicator of the criterion's validity. One indicator compared "the opinions on learning with didactic designs" which were made during the external tests as well as during the internal tests which were part of the AAS. Another indicator compared "the opinions made on both questionnaires regarding acquisition of computer skills."

Analysis of the validity of the AAS's Group Component

The collaborative phase focused on group production. In this section, the validation procedures related to AAS along with collective information are described.

Consistency among peer opinions on didactic designs. Once the groups created their didactic designs, the designs were submitted to two or three students that were not part of the group. Each one of these students independently evaluated his or her peers' work according to designed guidelines based on rubrics. The correlations between the evaluations of the different peers on the same design were used as a peer evaluation consistency index (I-5). The average of the indexes associated with evaluations of each one of the 8 designs created by the students was considered as the internal peer evaluation consistency index.

Consistency between the AAS evaluations on creation and validation of the didactic designs. The analysis, in this case, refers to the four evaluations regarding the creation and validation of the didactic designs (I-4 to I-7).

- *Consistency between the tutor, peer, expert, and student evaluations on didactic designs*. The first three evaluations (I-4, I-5 and I-6) were structured as part of a series of criteria defined by rubrics. Only the final evaluation, the self-evaluation (I-7), consisted in a grade, on a scale of 1 to 7, which corresponded to the student's overall appreciation of the design created by his or her work group.

The four evaluations, although they refer to the same object, i.e. to the didactic design, measured its diverse aspects at different times. Such measurements were made with different emphases, which were indicated by means of assigning different values to the items. As a criterion of consistency, the four valuations were correlated in two parts.

- *Consistency between the peer and the expert evaluation regarding the level of suitability of the activities for the learning expected to be achieved by means of the designs*. The didactic designs were assessed by the peers before they were applied in the classroom. Then, after they were applied (I-5), they were assessed by an academic didactics expert (I-6). The peers as well as the expert provided their opinions regarding the "suitability of the activities proposed in the designs to achieve the purpose of the designs themselves". The opinions refer both to the learning activities as well as to the evaluation activities. The opinions were made within the context of a scale associated with rubrics, and they made up part of AAS's instruments. In other words, the opinions made up part of the guidelines applied by the peers (I-5) and the guidelines applied by the academic didactics expert (I-6).

Web usability and participants' attitude as possible invalidation factors

A significant positive correlation between AAS results and factors such as web usability and participant attitude was understood as an invalidation factor of the assessment system.

Participants' attitude. The Attitude towards mathematics Test (Aikeen, 1996) was adapted to measure student attitude to the course's AAS. 24 items were organized according to a Likert scale. The test reliability was Cronbach α=0.62.

Usability. The Sanchez usability test (2000) was adapted to measure the usability of the AAS component of the web site. The 21 items referring to five factors called learning, satisfaction, error, efficiency, and memory were organized according to the Likert scale. The test reliability was Cronbach α=0.74.

Results

This section provides the test results used to characterize the AAS's validity and reliability. These tests mainly consisted in correlations as indicators of concurrent validity and coefficients of internal consistency; which were organized in four groups, namely: The tests that compared the AAS results with the student's learning and with the group productions in a real context; the isolated tests as indicators of validity of the AAS's individual component; the tests related to the validity of the AAS's group component; and, finally, the tests related to two possible invalidation factors.

AAS' relation with both real context products and individual learning

Consistency between AAS' results and products' effectiveness in real context. The averages of the children's results were compared with the teachers' AAS results. Two of the teachers' reports about design application did not provide enough information to be considered in the correlation analysis. As shown in table 5, some designs were implemented in more than one class. The Spearman correlation was positive, rho(6)=.30 p=.2, but not significant.

Table 5. Results of implemented designs

	Groups							
	Gr1	Gr2	Gr3	Gr4	Gr5	Gr6	Gr7	Gr8
Number of classes	3	2	2	1	4	m.v.*	3	m.v.*
Schoolchildren results	66%	80%	80%	98%	93%	m.v.*	74%	m.v.*
AAS average results	70	74	78	77	92	53	79	80

* missing value

Measurement of degree of association between AAS results and individual learning criteria. The Spearman correlation as a concurrent validity test gave a value *rho*=.24, p>0.05, which was not significant.

Validity of AAS's individual component

The individual component was instrumental instruction, meaning that, in this part of the course, they were given conceptual tools to elaborate didactical designs in order to progress to a collaborative approach during the next part of the course.

The students´ ability to recognize the components of a didactic design. The item *"*What are the main phases of a didactic design?*"* included on the pre-test, as well as on the online test (I-2) and on the post-test (E-3), had an unexpected variation regarding the number of correct answers in each of the applications: $Mean_{\text{pretest}}$ = 5/20 (25%), $Mean_{\text{on\_line\_test}}$ = 19/25 (76%), and $Mean_{\text{post-test}}$ = 12/21 (52%).

The online test and post-test results were expected to be consistent, particularly those items that were the same. Nevertheless, the correlations were not significant; they were low and even negative. The correlation was *rho*(19)=-.23, p>.05, (two-tailed).

The students' capacity to explore the website and send e-mails. The estimated level of association between external information and that gathered through AAS according to a Spearman correlation was *rho*(23)=0.44, *p*<0.05 (two-tailed). Upon including those subjects who did not answer some of the items in the analyzed questions, this indicator was more significant, *rho*(23) = 0.68, p<0.001 (two-tailed). This was backed by the presumption that failure to reply or send an electronic message at the appropriate time was an indicator of not knowing how to do so.

In either of the two cases, it may be claimed that there was information that provided complimentary evidence of the validity of this component within the system's scope of authentic assessment.

Students' opinion on their learning achievement regarding didactic designs and computers. The opinions referring to the learning achievement on didactic designs, gathered from the AAS questionnaire (I-7) and from the external questionnaire (E-1), correlated significantly, *rho*(11)=0.55, *p*<.05, (one-tailed). The correlation referring to the learning perception on Internet use was *rho*(10)=0.54, *p*<.05 (one-tailed).

Validity of AAS´s productive evaluation component

The group component was focused on a productive approach, where students were asked to build and apply a design.

Peer evaluation (I-5). The peer evaluations of the same design were quite coincidental. A high average of internal consistency for an average of 8 groups was attained: $\alpha(6)=0.96$, $p<.001$. This means that the evaluators had quite similar assessment criteria.

Consistency between AAS's group component evaluations. In this section consistency between the group evaluations in regard to the creation and validation of the didactic designs was analyzed.

- Consistency between the tutor, peer, expert, and student evaluations regarding the didactic designs (I-4 to I-7). By assigning points in terms of rubrics, the evaluators coincided that the designs created by the students met course specifications. This was reflected in the high scores. The groups averaged 94% when evaluated by the tutor, 82% when evaluated by their peers, 75% when evaluated by an expert, and 97% when self-evaluated.

Table 6. The four evaluations referred to the didactic designs

Evaluators	Groups								
	Gr1	Gr2	Gr3	Gr4	Gr5	Gr6	Gr7	Gr8	Average
Tutor	100%	70%	100%	100%	100%	80%	100%	100%	94%
Peers	72%	88%	92%	68%	76%	84%	84%	96%	82.4%
Expert	57%	80%	71%	88%	100%	29%	97%	74%	75%
Students	97%	93%	100%	100%	100%	93%	100%	93%	97%

The consistency between the tutor, peer, expert, and student evaluation results is expressed by correlations in Table 7. It is interesting to note that the correlation between the peer evaluation and the others was negative. It is even more interesting that the evaluations of two or more peers on the same didactic design were extraordinarily consistent, $\alpha=0.96$, $p<.001$. Furthermore, the tutor, expert, and student evaluations were consistent among themselves. All were positive and two of the three were significant, as shown in table 7.

Table 7. Correlation matrix referred to peers, expert, students, and tutor evaluations

	Expert	Student	Tutor
Peers	-.14	-.48	-.24
Expert		59*	.39
Student			.70**

- *Consistency between peer and expert evaluations*. The degree of agreement between peer evaluation and expert evaluation on the items referred to as "consistency between learning activities and anticipated learning" and "consistency between the evaluation activities and expected learning" can be seen in the table 8 and 9.

Table 8. Evaluation of the relationship between design and anticipated learning activities. By peers and by expert

Evaluators	Groups								
	Gr1	Gr2	Gr3	Gr4	Gr5	Gr6	Gr7	Gr8	Average
Peers	100%	100%	100%	50%	68%	83%	100%	50%	81.3%
Expert	100%	100%	-	100%	100%	100%	50%	50%	73.5%

Table 9. Evaluation of the relationship between evaluation and anticipated learning activities, by peers, and by expert

Evaluators	Groups								
	Gr1	Gr2	Gr3	Gr4	Gr5	Gr6	Gr7	Gr8	Average
Peers (4)	100%	90%	75%	75%	68%	83%	100%	50%	80.0%
Expert (8)	100%	100%	100%	100%	100%	50%	100%	50%	87.5%

The correlations were not high enough to be significant. They were *rho*(5)=0.11, p>.05 and *rho*(6)=0.46, p>.05, respectively. As one can observe, both the expert and the peers assigned high scores to their evaluations. The slight differences between them led to null correlations.

Web usability and participants' attitude as possible factors of invalidation

Attitude of participants. Students' attitudes to the AAS aspects of the WEB were positive, and the association of this variable to AAS results was null, rho(20) = 0.09.

Web usability. Students considered AAS' aspects of the web usability to be adequate. The correlation between students' evaluation of AAS aspects of web usability with their results in AAS was positive, but not significant. The Spearman correlation was rho(15)=0.27. This means there was no significant relation to affirm that the variables were directly related.

Discussion

This chapter summarizes the study results, analyzes the technical problems of the AAS, and offers suggestions to elaborate an Authentic Assessment System for a web-based course.

Study results: Validity of AAS

This study considered three perspectives to analyze the Authentic Assessment System's validity. The first perspective, using concurrent validation techniques, analyzed the AAS's validity as a whole, as a one-dimensional variable. The second perspective, also using correlation techniques, analyzed local aspects of the AAS, focusing on performance in real contexts and non-obstructive data. The third perspective was a "reciprocal" modality; instead of validating the AAS according to its degree of association with other criteria, the objective was to demonstrate a null association between the AAS and possible invalidation factors of the system.

One-dimensional perspective. From this view, the AAS results were considered as whole values, which were correlated with two external variables. One was the degree of effectiveness of the didactic designs elaborated by the students, and the other was the students' results in a posttest. The two correlation indexes were positive, but neither was significant; thereby confirming the difficulty in validating the authentic assessment procedures, as several researchers warned (Linn et al., 1991; Aiken, 1996; Herrington and Herrington, 1998; Schurr, 1999). The question remains if other indexes inform some degree of validity of the AAS used in the course.

Desegregated perspective. Considering that the AAS was multidimensional and that diverse factors affected the results, this second approach separately analyzed the AAS's group component from the individual component. These analyses were limited to selected data, and some gathered non-obstructively. Out of the five indicators used to analyze the individual component, four correlated significantly: the ability to search in the web, the ability to use e-mail, the opinion about ICT learning, and the ability to build didactic designs. On the other hand, only three indicators used to analyze the AAS's group component were significantly favorable, which can be partially explained by the small number of student groups. In fact, the third indicator, in spite of reaching the correlation rho = 0.46, was not significant. This second analysis perspective was fruitful, offering some insight in terms of the characteristics of an authentic assessment system that evidences validity.

Reciprocal perspective. The third perspective of analysis of the AAS's validity confirmed that we cannot attribute the students' success to either of the two invalidation factors considered, the attitude towards the evaluation system and its usability. Under this perspective, the AAS offered an additional partial evidence of validity.

Then, the results of this exploratory correlational case study indicated some evidence of validity, but also several technical problems with the authentic assessment system. We may conclude that the authentic assessment system implemented with two components, one individual and the other in groups, showed some problems and some insights.

The following sections offer interpretations based on the in-depth analysis of available data. These interpretations give rise to suggestions for elaborating authentic assessment systems and to formulate hypotheses in subsequent studies.

Reliability and validity of each AAS's component

AAS's individual component. It should be mentioned that the correlation between AAS's results and the posttest achievement was not significant. Although AAS's results were good with 89% effectiveness, they cannot be understood as learning in the manner that the posttest measured it. Even though the post-test pre-test differences were significant according to the t-test ($p<0.01$, $n=18$), this learning indicator was lower than what was expected. In fact, with an initial achievement level of 25% on the pretest, only an average level of achievement of 48% was attained on the posttest. These results differed greatly from the individual achievements on the AAS (89%). There was a particularly large difference with the online test taken two months earlier, where average performance was 78%.

Curiously, the online test results correlated with the number of student attempts to submit them ($rho=0.66$, $p<0.01$), but they did not correlate with the posttest results. On the other hand, there was consistent evidence between external information and that gathered through AAS, which referred to the students' capacity to explore the website and send e-mails and the students´ opinion on their learning achievement. Therefore, although specific evidence of the validity of AAS individual learning existed; as whole instruments, both the online test and the posttest were weak, revealing reliability and validity problems in this individual component.

AAS´s group component focused on group production. Although designs were effective and AAS results were good, their correlation was not significant. The tutor, expert, and student evaluations were consistent among themselves. However, correlations between the peer evaluation and the other evaluations were null, even negative. Surprisingly, the peer evaluations on the same design were coincidental; both the expert and the peers assigned high scores to the designs, and the slight differences between them led to null and negative correlations. Group six's final report (see Table 9) provided the information source that led to incongruence between the peer evaluation and the expert evaluation. The report was incomplete, which made an impact on the criterion adopted by the expert upon evaluating the items in question. Thus, we can presume that some inconsistencies came from the data, which was collected from real context in non-experimental designs. Other inconsistencies came from the instruments that measured multidimensional factors based on criteria, and non-standardized evaluations during the learning process.

Relationship between the individual and group components

Upon viewing the course's individual component as preparation for the eminently productive group component, no causative relationship could be perceived. The products attained through collaborative work were shown as the result of an interaction in which the individual contributions were not clear.

The group evaluations were connected with the creation and validation of the didactic designs. Even though they made up 75% of each student's grade, they could not be understood as evaluations of individual learning in the traditional sense of the term. From this perspective, the low correlation between the authentic assessment results and the post-test's, an individual multiple-choice test, revealed this discrepancy. It does not constitute a source of AAS invalidation.

The scarce evidence of AAS validity did not lie in the low association between these measurements, but rather in the low indices of consistency in those aspects that were expected to be high. In spite of this, there was a reasonable consistency index between them.

First suggestion: Seek a balance between reliability and validity

The first suggestion of this work is to "Seek a balance between reliability and validity" when you are building a Web based course with an AAS. The suggestion arises from the arguments covered in this section.

Data illustrate some reliability and validity problems, which are very typical of authentic assessment. Because effectiveness of products reached through the course referred to variables measured in context, it was difficult to obtain a good correlation with both the process and the results of the course. Similarly, since authentic assessment is related to complex learning in special multiple dimensions, it was also difficult to obtain a good correlation with a single multiple-choice test.

We should mention that AAS instruments were technically designed to measure knowledge, abilities, competency, and attitudes in accordance with the main principles of authentic assessment. In addition, AAS has evolved into a highly effective course. So in spite of the evidence of lack of reliability, this assessment system shows evidence of cognitive complexity, significance, and efficiency; that is, authentic validity.

Normal-based and criteria-based evaluation. According to an "edumetric" assessment model, almost all people are expected to learn, but according to the psychometric measurement model, normal distribution is expected to describe individual variability in learning. In this second view, learning differences between higher and lower groups are expected to be stable and significant. So, reliability is expected as a necessary condition of validity. According to the former view, significant differences are not expected between learners. So, some slight differences arising from uncontrolled variables are expected. Therefore, stability of differences in results should not be considered as the central point of an effective course validation.

Straetmans et al., (2003) assert that principles of classical test theory may be applied to qualitative assessment, and measurements of quality control such as validity and reliability should be applied to forms of competency assessment in a similar way as they were used for other tests. We should keep in mind that psychometric refers to maximizing differences between individuals and edumetric refers to measuring within-individual growth, without reference to other individuals. Frederiksen & Collins (1989) definitively propose specific criteria, and criteria other than validity, for newer forms of assessment.

Valid measures with low reliability. Several correlations used in this study refer to the consistent association between variables, in relation to reliability. Stability, consistency, and concurrent validity were not reached in data. How much stability should be expected as part of the validation process? Can valid measures with low reliability be obtained, as is often the case in authentic assessment systems and was this the case in this study? We understand that both reliability and validity are expected in an assessment process. But, following the traditional path, validity is sacrificed in order to obtain reliability, and according to an authentic assessment approach, soundness of measurement is privileged as compared to stability of results.

Four major problems that affect AAS validation

Online tests, rubrics measuring, self-evaluation techniques, student dropout, and restrictions in timing tasks, which are common in distance education, demonstrated some technical problems. In effect, the online tests with the possibility of repeating answers is questionable; the self-evaluations should be adjusted to more specific referents; the use of rubrics looks promising but needs to be improved; and the various participants' contributions to the evaluation process, which has proven to be effective and widely accepted, still needs to be adjusted.

Validity of the online test with multiple attempts. The opportunity granted to the students to answer the online test after several attempts did not favor AAS's reliability. The online test was part of the authentic assessment system's individual phase. In order to complete the test, the computer system allowed the students to make several attempts before recording their final answers. Even though this favored AAS performance, the online test was not confirmed as being a contributing factor to the improvement of post-test scores. Students seem to only partially retain the knowledge that was required of them on the posttest, which they had taken two months after taking the online test.

The opportunity provided by the online test proved to be attractive to the students. Once the test was completed, the computer system automatically sent them the test results with percentages for each group of questions. Since these percentages referred to groups of items and not each individual item, and the computer system assigned partial credit to the partially correct answers, it was not easy for the students to know which items they had gotten wrong. They needed to analyze each item of a group in order to improve their test scores.

This opportunity may have had a positive short-term effect without requiring long-term retention of the material. The knowledge being tested on the online test was part of the theoretical knowledge that the students would use when creating their didactic designs. Therefore, we expected that this theoretical knowledge would be meaningful to them, thus remaining in their long-term memory until taking the posttest. Apparently thought processes during the on line test were not sufficient to consolidate such knowledge.

Validity of AAS's group component: This component included tutor, peer, expert and student evaluations. Only partial evidence of the consistency between the tutor, peer, expert, and student evaluations regarding the creation and validation of the didactic designs was obtained. The peer and expert evaluations agreed that the didactic designs were good. However, they were not coincidental enough when establishing a hierarchy with regard to that quality. Although the evaluations carried out by the different participants were positive and the didactic designs in question were effective, the evaluations were not consistent enough among themselves. Why was there not enough consistency between the peer evaluations and the expert evaluation in the cases when they assessed the same aspect? The dissonance could have its origins in the assessment instruments' deficiencies and in the research design's insufficiency. This would not have allowed the assimilation of the fact that the students' behavior varied over time.

The group component evaluation that was not carried out on the basis of rubrics was a self-evaluation of the students. The assessment consisted of grading the didactic design on which they worked. Apparently, it is necessary to check the effect of subjectivity on the self-evaluations. We suggest doing this because the students evaluated their designs quite positively. On average, they assigned themselves a score of 97%. This differed from the expert's evaluations of the reports on didactic designs, which averaged only 75%. The objectivity might be increased if the students also assigned themselves a grade on the basis of rubrics and not based on holistic perception.

Dropouts. Another element to consider is the number of groups about which the data are gathered. The low number of only eight work groups constituted a limitation on the analysis of AAS's group component. Statistical significance is difficult to obtain with a limited number of cases. In addition to following up on evaluation procedures and the methodological design adopted to examine AAS validity, an analysis of the students' behavior should be conducted. Not all of the students finished the course as expected. Only 52% (13 of 25) of the students submitted their self-evaluations despite the fact that it had a detrimental effect on their grade. Only five of the eight groups submitted a completed final report. In spite of these deficiencies, the students were able to pass the course with a good grade. The students would have liked to have finished the course by submitting all of their work. However, the due dates expired and, faced with high demands at their jobs, they did not meet the final requirement.

Mortality, referring to subjects who were not included in the analysis because they did not provide certain responses, had an important effect on the study. In fact, if the omitted answers had been considered incorrect and their corresponding instruments had received a score of zero, not only would we have obtained positive correlations, but also the majority of them would have been significant. This would have corroborated the validity of the AAS.

Time factor limits production. During the creation of the didactic designs, the evaluators indicated the weak aspects of each didactic design. This gave the students the opportunity to correct their designs before the next evaluator's revision. This allowed, within a highly limited time frame, time margins for some groups of students to improve their weakest aspects. Thus, the designs had the tendency to improve. However, the groups' work could not always be completed within the established period of time. This was noticeable at the end of the course, when the expert conducted the final evaluation of the didactic designs subsequent to their application in the classroom. The moment arrived when the students had to deliver their reports in whatever state they were in, and in some cases this implied an interruption in their production.

Final suggestions to improve evaluation systems in web based courses

We expect improving assessment instruments and the assessment system design of a web based course by means of successive approximation processes: Multidimensional authentic assessment demands, integrated techniques, purified instruments, and an extensive sample to achieve acceptable reliability.

Refining and integrating evaluation instruments. Using integrated evaluation procedures that pervade the course's main concept seems to be appropriate. Considering that the authentic assessment system is multidimensional by nature, some specific core objectives would be identified, and the number of variables to be assessed should be reduced. Thus, the evaluation process would be permanently focused on those objectives with the aim of monitoring progress~~.~~.

Bearing in mind that the results of the self-evaluations and the evaluation through rubrics were very elevated, the question arises whether those results came from low requirement levels. In order to supply a satisfactory answer, standardized evaluation instruments are vital, which unfortunately stray from the authentic assessment method.

Self-regulation techniques such as self and co-evaluation proved to be attractive, because they fostered course effectiveness. However, they need to be refined in order to attain better reliability indices and evidence of validity.

Self-regulation techniques seem to be a high priority; experts establish more rigorous criteria on the quality of the products to be created by students, safeguarding AAS's ecological validity. In a course for teachers, didactic designs effective in the classroom should be supported on the basis of testing the schoolchildren on whom the designs are applied, with validity recognized by peers as well as by the expert. AAS should provide the guarantee that classroom achievements are adjusted to anticipated learning according to the official attainments.

Even though pertinence and effectiveness of didactic designs are essential requirements for favoring the assessment system validity, they are difficult to meet.

In this study, the validation of AAS's group component provided evidence that the procedures and instruments employed require greater integration, as the literature recommends (Chong, 1998; Roschelle, 1992; Lave, 1991).

Improving Rubrics: Rubrics need to be further developed. Their use requires a great deal of specificity in their operation. Perhaps it is appropriate for the evaluators to limit themselves to areas of specific domain. An example would be for the same evaluator to judge the same aspect in several didactic designs instead of judging all of the dimensions of a specific design.

Improving the authentic assessment design. When assessment design includes both individual and group components, a differentiated evaluation could be instated within each group on the basis of individual roles. The students could assume differentiated tasks that are assessed individually even though the final product is a shared responsibility. For example, a student can act as a graphic designer and learn more than his or her peers about the use of related software. In each subject, individual growth, according to previous experience, expectations, and variety of interactions generated within the group, can be verified. In this manner, we assume a priori that the types of learning are different. However, at the same time, they support each other and enrich the collaborative work's productivity.

Even though this approach would be far from the traditional approach that measures all students with the same stick, it would be consistent with the principles of an authentic assessment and would favor AAS's ecological validity.

A research design that takes recommended variations into account should be assimilated into a time series analysis. The data should be gathered gradually throughout the process so that progress may be observed and consistency between measurements may be appreciated. Within that context, information should be gathered about those aspects of the students and their work in which change is expected, which at the same time provides feedback to the individual learning and group work processes.

Qualitative data support

Qualitative data were gathered during the AAS application process. Content of students' e-mails, an open-ended questionnaire, two focus group discussions, and notes from participant observation strengthened our understanding of the AAS limitations. A qualitative approach was not developed in these pages because of its length; however some commentaries were provided that appear in the discussion section. These comments are specifically about online test multiple attempts and the reasons why some students did not meet the final course requirements (see discussion points) Absolute qualitative data enrich our understanding and improve AAS information for subsequent researchers to use for new studies.

Summary

This section summarizes the study's results and discusses the weaknesses in the validity and reliability of the AAS, which emerge from the contextual and multidimensional characteristics of the evaluation in a non-experimental design. Then it discusses the importance of achieving a balance between the reliability and validity criteria in an Authentic Assessment System.

The following section associates the weaknesses of the AAS to the Online Test modality, to the use of rubrics and self-evaluation and co-evaluation techniques, to the drop out rate, and to the time restrictions of an evaluation in context. The thoughts about these weaknesses give rise to suggestions for improving assessment systems in web based courses.

Finally, suggestions are given for improving the validity and reliability of the AAS in a gradual process using time series design and maintaining the measurements in real contexts. Improvement was focused on the evaluation instruments and in the system's design. The suggestions regarding the instruments were: to use integrated instruments, refined rubrics, and rigorous quality criteria. The suggestions for improving the design of the authentic assessment are: to reduce the number of specific objectives, to harmonize the individual component with the group component, if applicable, to include self-regulation procedures, to apply differentiated evaluation in accordance with the roles in the collaborative work and to the potentialities of the peer evaluators and/or experts, in order to favor a context of effectiveness and ecological validation.

References

Aiken, L. R. (1996). *Tests Psicológicos y Evaluación*, México. Prentice Hall.

Bandura, A., & Walters, R.H. (1963). *Social Learning and Personality Development*, New York: Holt, Rinehart and Winston.

Benzie, D. (1999). Formative evaluation: Can models help us to shape innovative programmes? *Education and Information Technologies, 4* (3), 251-262.

Chang, C.-C. (2002). Assessing and Analyzing the Effects of WBLP on Learning Processes and Achievements: Using the Electronic Portfolio for Authentic Assessment on University Students' Learning. *Paper presented at the EdMedia 2002 Conference*, June 24-29, 2002, Denver, CO, USA.

Chong, S.M. (1998). Models of Asynchronous Computer Conferencing for Collaborative Learning in Large College Classes. In C. J. Bonk & K. S. King (Eds.), *Electronic collaborators: learner-centered technologies for literacy, apprenticeship, and discourse.* Mahwah, N.J.: Lawrence Erlbaum Associates, 157-182.

Clarke, M., Butler, C., & Schmidt-Hansen, P. (2004). Quality assurance for distance learning: A case study at Brunel University. *British Journal of Educational Technology, 35* (1), 5-11.

Collis, B., De Boer, W., & Slotman, K. (2001). Feedback for web-based assignments. *Journal of Computer Assisted Learning, 17* (3), 306-313.

Condemarin, M., & Medina, A. (2000). *Evaluación Auténtica de los Aprendizajes: Un medio para mejorar las competencias en lenguaje y comunicación*, Santiago de Chile: Editorial Andrés Bello.

Downs, E., Carlson, R. D., Repman, J., & Clark, K. (1999). Web-Based Instruction: Focus on Learning. Georgia Southern University. *Paper presented at the SITE Conference*, February 28-March 4, 1999, San Antonio, TX, USA.

Duart J., & Sagrá, A. (2000). *La formación en web: del mito al análisis de la realidad*, retreived October 15, 2007, from http://cvc.cervantes.es/obref/formacion_virtual/campus_virtual/sangra2.htm.

Fenwick, T., & Parsons, J. (1998). Starting with our stories: Towards more authentic assessment in adult education. *Adult Learning, 26* (4), 25-30.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*, 27-32.

Gatlin, L., & Jacob, S. (2002). Standards-Based Digital Portfolios: A Component of Authentic Assessment for Preservice Teachers. *Action in Teacher Education, 23* (4), 28-34.

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*, London, The Falmer Press.

Herman, J., Aschbacher, P., & Winters, L. (1997). *A practical guide to alternative assessment*, CA: CRESST.

Herrington, J., & Herrington, A. (1998). How university students respond to a model of authentic assessment. *Higher Education Research and Development*, *17* (3), 305-322.

Kroll, D., Masingila, O., & Mau, S. (1992). Grading Cooperative Problem Solving. *The Mathematics Teacher, 85* (8) 617-626.

Lajoie, S. (1991). A framework for authentic assessment in mathematics. *NCRMSE Research Review, 1* (1), 6-12.

Lave, J. (1991). Situating Learning in Communities of Practice. In L. B. Resnick, J. M. Levine and S. D. Teasley (Eds.), *Perspectives on socially shared cognition*, Washington, DC: American Psychological Association, 63-82.

Linn, R. L., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20* (8), 15-21.

Meyer, C.A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership*. *49* (8). 39-40.

National Center for Fair and Open Testing (1992). *What Is Authentic Assessment?* Cambridge, MA. NCFOT.

Newmann, F., & Wehlage, G. (1993). Five Standards of Authentic Instruction. *Educational Leadership, 50* (7), 8-12.

Olson, T., & Wisher, R. (2002). The Effectiveness of Web-Based Instruction: An Initial Inquiry. *International Review of Research in Open and Distance Learning*; *3* (2), retrieved October 15, 2007, from http://www.irrodl.org/index.php/irrodl/article/view/103/182.

Orde, B. (2001). Online course development: summative reflections. *International Journal of Instructional Media, 28* (4), 397-403.

Peterson, M. (2000). Electronic Delivery of Career Development University Courses. In J.W. Bloom, & G. R. Walz (Eds.), *Cybercounseling and Cyberlearning: Strategies and Resources for the Millennium*, Alexandria, VA: American Counseling Association, 143-159.

Phipps, R., & Merisotis, J. (1999). *What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education*, Washington DC: The Institute for Higher Education Policy.

Reeves, T. C., & Okey, J. (1996). Alternative assessment for constructivist learning environments. In B. Wilson, (Ed.), *Constructivist learning environments*, Englewood Cliffs, NJ: Educational Technology, 191-202.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gilford & M.C. O´Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction*, Boston: Kluwer, 37-75.

Roschelle, J. (1992). What Should Collaborative Technology Be? A Perspective from Dewey and Situated Learning. *ACM SIGCUE Outlook, 21* (3), 39-42.

Ryan, W. J. (2001). *Comparison of Student Performance and Attitude in a Lecture Class to Student Performance and Attitude in a Telecourse and a Web-Based Class*, Ph.D. Dissertation No. ED467394, Nova Southeastern University.

Salomon, G. (1992). What Does the Design of Effective CSCL Require and How Do We Study Its Effects? *ACM SIGCUE Outlook, 21* (3), 62-68.

Scardamalia, M., & Bereiter, C. (1996). Computer Support for Knowledge-Building Communities. In T. Koschmann (Ed.), *CSCL, theory and practice of an emerging paradigm*, Mahwah, NJ: Lawrence Erlbaum Associates, 249-268.

Schurr, S. (1999). *Authentic Assessment From A to Z*, USA: National Middle School Association.

Stake, R. (1995). *The art of case research*, Thousand Oaks, CA: Sage.

Straetmans, G., Sluijsmans, D., Bolhuis, B., & Van Merrienboer, J. (2003). Integratie van instructie en assessment in competentiegericht onderwijs [Integration of instruction and assessment in competency-based education]. *Tijdschrift voor Hoger Onderwijs, 21*, 171-198.

Torrance, H. (1995). *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment*, Buckingham: Open University Press.

Tucker, S. (2000). Assessing the Effectiveness of Distance Education versus Traditional On-Campus Education. *Paper presented at the Annual Meeting of the AERA*, April 24-27, 2000, New Orleans, Louisiana, USA.

Vygotsky, L. S. (1985). *Thought and Language*, Cambridge, MA: MIT Press.

Weiss, C. (1997). *Investigación Evaluativa. Métodos para determinar la eficiencia de los programas de acción* (4th Ed.), Mexico: Editorial Trillas.

White, S. (1999). The effectiveness of web-based instruction: A case study. *Paper presented at the Annual Meeting of the Central States Communication Association*, April, St. Louis, MO, USA.

Wiggins, G. (1989). A True Test: Toward More Authentic and Equitable Assessment. *Phi Delta Kappan, 70* (9), 703-713.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments to Inform and Improve Student Performance*, San Francisco: Jossey-Bass.

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, *2* (2), retrieved October 15, 2007, from http://pareonline.net/getvn.asp?v=2&n=2.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *75* (3), 200-208.

Wilkerson, J., & Elkins, S. (2000). CAD/CAM at a Distance: Assessing the Effectiveness of Web-Based Instruction to Meet Workforce Development Needs. *Paper presented at the Annual Forum of the Association for Institutional Research*, May 21-24, 2000, Cincinnati, OH, USA.

Woolf, B. P., & Regian, J. W. (2000). Knowledge-based training systems and the engineering of instruction. In S. Tobias and J. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military*, New York: Macmillan Reference, 339-356.

Weller M. (2002). Assessment Issues on a Web-based Course. *Assessment & Evaluation in Higher Education, 27* (2), 109-116.

Yin, R. (1984). *Case study research: Design and methods* (1st Ed.), Beverly Hills, CA: Sage.

Yin, R. (1994). *Case study research: Design and methods* (2nd Ed.), Beverly Hills, CA: Sage.

Young, M.F. (1995). Assessment of situated learning using computer environments. *Journal of Science Education and Technology, 4* (19), 89-96.