The Authenticity of Authentic Assessment: What the Research Says...Or Doesn't Say

Brian McAlister
University of Wisconsin-Stout

The purpose of this chapter is to report what the research says about authentic assessment. First, the claims that have been made about the benefits of authentic assessment as a mechanism for measuring student performance are discussed. Next, the claims that have been made about the benefits of authentic assessment as a mechanism for facilitating learning are examined, followed by a review of research related to authentic assessment. The chapter concludes with a brief discussion of some key issues of concern related to research and practice.

Much has been written about the promise of authentic assessment. A primary focus of much of what has been written is to promote the use of authentic assessment as a superior alternative to other forms of assessment. Although much of this work has been positive and while many of the benefits make sense intuitively, the question remains: What does the research say about the benefits and problems associated with authentic assessment? To understand some of these claims, it is important to understand the conceptual foundations of authentic assessment.

Authenticity and Authentic Assessment

The overarching theme of authentic assessment is, as the term indicates, authenticity. This thrust relates both to the authenticity of the learning activity as well as the authenticity of the assessment. One concern that is voiced throughout the literature has to do with what makes an activity “authentic.” In vocational education circles, with the rich history of laboratory-based learning, this concern is much less problematic than in the more traditional academic areas. As performance and authentic assessment have moved more broadly into the academic arena and as vocational and academic education have attempted to work more closely together, the issues of “authenticity” have become more important.

Messick (1992) captures this sense, indicating that “a fundamental ambiguity pervades authentic educational assessments, namely, authentic to what?” (p. 27). He poses the question of whether assessments should be authentic reflections of classroom work or authentic reflections of the “real world.” This is a subtle, but important distinction. What is meant by the real world? Sometimes students are taught using conventions that have been found to be effective and efficient
methods of teaching certain skills or concepts. Does the fact that educators commonly use various algorithms when teaching mathematics make this approach authentic? Is authenticity defined by the boundaries of the classroom? To what extent do (and should) the real world experiences of students coincide with what occurs in classrooms and laboratories? Students’ perceptions of the real world may indeed be very different than those of their teachers. One could assume that authenticity means teaching in context or contextual learning. But then we are left to ascertain which of these contexts are worthy of the distinction of being considered “real” or “authentic.” For example, some vocational schools have established automotive service programs that operate like service centers in automotive dealerships. Customers schedule their maintenance with students, who order parts, repair the automobiles, and when the work is completed, bill the customers. Because this is an educational experience, there are times when an instructor must intervene on behalf of the customer. Students cannot be allowed to make serious mistakes that could result in a dangerous automobile being released to the customer. This scenario poses some serious questions. If authentic assessments should reflect the “real world,” How real is “real”?, Whose reality should it reflect?, and What degree of authenticity is “authentic”? It is obvious from this example that educators must apply reasonable limits on authenticity as a function of concerns such as safety, confidentiality, and more.

Others have attempted to clarify these issues by suggesting criteria to gauge the authenticity of an activity or learning experience. Newmann and Wehlage (1993) suggest that, in order for instruction to be considered “authentic,” students must construct meaning and produce knowledge, use disciplined inquiry to construct meaning, and aim their work toward production of discourse, products, and performances to a level of value or meaning beyond success in school.

In order to meet these criteria, Newmann and Wehlage offer five standards (or criteria) that can be used to distinguish levels of authenticity of a learning activity:

1. To what extent are students required to use higher-order thinking skills?
2. What is the depth of student knowledge and understanding that is attained?
3. At what level does a learning or assessment activity have value and meaning beyond the classroom?
4. To what extent are students required to discuss, learn, and understand the substance of a subject?
5. How well does an assessment measure the expectations, respect, and extent of inclusion of all students in the learning process?

Newmann and Wehlage’s criteria are useful because they refine and clarify the distinctions that should be made relative to the meaning of authenticity. The criteria also extend authenticity beyond simple participation in “real” experiences to active reflection on the meaning of those experiences.

From another perspective, Cronin (1993) and Tanner (1997) suggest that the concept of authenticity is relative and exists along a continuum. An example of
this would be to compare activities that might occur in a teacher education program. It makes sense that demonstrating how to use a cooperative learning technique during a microteaching activity could be considered more authentic than simply writing a paper about cooperative learning. On the other hand, using cooperative learning techniques in a class while student teaching would be considered more authentic than using the same techniques during microteaching. Drawing from this example, it is apparent that learning activities can be placed along an authenticity continuum. Cronin (1993) supports this approach by suggesting that learning activities are “neither completely authentic nor divorced from reality” (p. 78). He further suggests that our goal as educators should be to move instruction toward the more authentic end of this continuum.

Another key aspect of authenticity in assessment relates to the strategy or system that is used. In order for assessment to be considered authentic, there should be consistency between the assessment and the real-world application for which the learner is being prepared (Tanner 1997). For example, if students are expected to be able to troubleshoot the electrical system of an automobile, then the assessment strategy should be designed in such a way as to be able to tell whether they have the knowledge and skill to perform that kind of activity.

Messick (1992, 1996) has analyzed the appropriateness of using authenticity as a standard for validity in assessment. He frames the issues in terms of representation, directness, and relevance. An assessment that suffers from construct underrepresentation variance fails to test a construct adequately, because a major aspect of the construct extends beyond the measure. For example, an assessment could be designed to measure whether a student can service automobile braking systems. If, however, students are only tested on one type of braking system (e.g., disk brakes), then this assessment would suffer from construct underrepresentation. “The measurement concern of authenticity is that nothing important has been left out of the assessment of the focal construct” (Messick 1996, p. 16). An assessment that suffers from construct irrelevant variance includes information that is irrelevant to the construct being tested. For example, the purpose of the assessment could be to determine whether students can apply appropriate design principles when designing a visual message. If the assessment method is restricted to identifying the parts of a camera, the assessment would suffer from construct irrelevant variance. Thus, an assessment is considered representative when it is broad enough to assess adequately the constructs being tested and direct when it is narrow enough to not be confounded with irrelevant information. Wiggins (1993) summarized a similar point, indicating that “tests are simplified of contextual ‘noise’ and ‘surround’ to make scores more reliable. Yet we need to maximize the fidelity and comprehensiveness of the simulation for validity reasons” (p. 230).

Tanner (1997) provides a good summary of the interrelationship between authenticity and learning experiences noting that——

[Authentic assessment] presumes that students will produce something that reflects not a narrow, compartmentalized repetition of what was presented to them, but an integrated scholarship which connects their
learning housed in other disciplines and which is presented in a setting consistent with that in which the learning is likely to be most useful in the future. (p. 14)

**Psychometric Issues**

Some disagreement exists in the literature regarding what sort of standards should be used to gauge authentic assessment from a psychometric perspective. Hipps (1993) argues that the assumptions underlying authentic assessment have their basis in constructivist theory. These assumptions, and associated psychometric considerations, are different from those commonly associated with traditional measurement theory. Therefore, he calls for a new set of standards that he suggests should start with trustworthiness and authenticity to replace the traditional standards such as reliability, validity, and objectivity, which are used in positivistic, quantitative research.

Reckase (1997) counters that this call for a different theoretical framework makes sense “if performance assessments are used solely as instructional tasks” (p. 12). However, if the issue is assessment then some statistical requirements are needed. He goes on to argue that “reasonable statistical requirements for sound performance assessments can be described based on current experience in the areas of (a) rater reliability, (b) test reliability, (c) generalizability, and (d) validity” (p. 3).

In a similar vein, Messick (1992) argues that in authentic assessment “different psychometric models might be employed . . . but such basic assessment issues as validity, reliability, comparability, and fairness still need to be uniformly addressed” (p. 7). He argues that “the interpretation and use of performance assessment . . . should be validated in terms of content, substantive, structural, external, generalizability, and consequential aspects of construct validity. These general validity criteria can be specialized for apt application to performance assessment, if need be, but none should be ignored” (p. 41).

One of the difficulties associated with understanding authentic assessment conceptually stems from the breadth of the assessment approaches that are currently being implemented, as well as the similarity among some of the terms. The issues are both substantive and rhetorical. Substantive issues have to do with such matters as psychometric practice, qualitative/quantitative distinctions, and the relationship between learning and assessment. At the rhetorical level, there is a general lack of precision related to what has become an almost interchangeable use of terms such as authentic, alternative, and performance assessment. Considerable work remains to be done to clarify the conceptual and practical distinctions among these terms (and associated practices).

Another factor that militates against gaining a better understanding of authentic assessment is that not all of the approaches that are being used can be categorized exclusively into discrete categories. For example, portfolios have been promoted as one viable method for making assessment more authentic. But all portfolios are not designed to document authentic learning activities. It is quite possible for
portfolios to contain relatively little that could be classified as authentic. In reality, portfolios typically contain a mixture of authentic and traditional assessment materials. In addition, there are many methods and tools used during the assessment process that cut across assessment categories, such as rubrics, observations, and self- and peer evaluations. But it is important to understand that just because a scoring tool such as a rubric is applied, the assessment is not automatically authentic. The key is to place the emphasis on the authenticity of the activity and whether the assessment strategy appropriately reflects the ability of students to apply what they have learned outside the classroom.

In summary, authentic assessment can involve a mixture of authentic learning and authentic assessment experiences. The first step is to develop activities that require students to apply, integrate, and synthesize knowledge and skill in a manner that reflects the real world and transcends the classroom. Apprenticeships and work study programs are exemplars of approaches vocational educators have used that are set in authentic learning environments. Similarly, it is also expected that assessment strategies should reflect the real world and that they should align with instructional goals and learning experiences. Authentic assessment experiences should, to the extent possible, not be contrived and will often involve multiple measures across time to provide a comprehensive picture of students' knowledge and abilities. It is best to conceive of authenticity as a continuum, representing activities that are totally contrived at one end to those that reflect the real world on the other.

There is currently some disagreement in the literature regarding what sort of standards should be used to gauge authentic assessment. Whereas Hipps (1993) calls for a new set of standards based on constructivist learning theory, Reckase (1997) and Messick (1992, 1996) support the need to retain, and perhaps refine and recast, traditional measurement standards such as validity and reliability. Although this issue is still up for debate, measurement standards, when reported in the research, are predominately discussed in traditional measurement terms.

**Authentic Assessment—The Claims**

Proponents of authentic assessment have made a variety of claims. Most of these claims fall within two broad categories: improved assessment and improved learning. These are addressed in turn.

**Authentic Assessment as a Means of Assessment**

It is difficult to discuss alternative assessment without using traditional assessment approaches as a frame of reference. Throughout the authentic assessment literature, there is a rather clear bias against traditional assessment approaches, which typically rely heavily on multiple-choice test items. This perception tends to be reinforced by the fact that nearly every state now mandates standardized testing (Henderson and Karr-Kidwell 1998), which relies heavily on such closed response test items. These tests are influencing educational practices because, in some instances, results are being used as indicators of teacher job performance and are
subsequently affecting teachers' salaries. Critics argue that this practice has resulted in a narrowing of the curricula, due to some teachers' resolve to teach to the test, thus corrupting the entire teaching-learning process (Henderson and Karr-Kidwell 1998; Shepard, Flexer, Hiebert, Marion, Mayfield, and Weston 1994).

Proponents of authentic assessment also worry that traditional forms of assessment (including tests and quizzes) fail to provide a holistic "picture" of student performance and knowledge over time. Traditional measures are designed to yield "snapshots" of what learners know at a given moment. To exacerbate the problem, many of the procedures used to prepare for these types of "snapshot" assessments tend to militate against learning transfer, synthesis, and retention (i.e., cramming and focusing on memorizing facts). These approaches typically do not engage students in authentic tasks and they tend to occur in an artificially contrived environment that does not reflect an activity they are likely to be called upon to do in the real world.

Another argument against traditional assessment practices is that there is an excessive emphasis on paper-and-pencil testing, which encourages the memorization of information. This results in higher test scores shortly following a lesson, while sacrificing long-term retention. Therefore, the goal of authentic assessment should be to provide a comprehensive, holistic, and robust "moving picture" of students' learning experiences by weaving assessment seamlessly into the teaching/learning process.

Most claims of improved assessment can be traced to the premise that if an assessment activity more closely resembles real-world practices, it must be more authentic and thus more valid. Simon and Gregg (1993) claim that "assessment becomes part of the instructional process, and vice versa, as planning evolves based on student progress toward goals, thus increasing the validity of such measures" (p. 4).

**The Impacts of Authentic Assessment on Learning**

Claims about the positive impacts of authentic assessment on teaching and learning are found throughout the literature. These are so common that it would be impossible to discuss them all in a single chapter. A few of the most common are discussed here.

One of the more general and pervasive premises is that learning experiences that reflect real-world activities are more valid. This validity represents more meaningful educational experiences that are proposed to be the driving force behind improved learning. "The expected positive effects of performance assessments on teaching and learning follow from their substantive validity" (Shepard et al. 1994, p. 6).

Another claim made by both researchers and educators is that authentic assessment experiences can improve student learning (Darling-Hammond, Ancess, and
Falk 1995; Shepard et al. 1994). Many of these claims are closely associated with a constructivist view of knowledge generation. The California Assessment Collaborative (1993) suggests that authentic assessment activities engage students in instructional tasks that require them to construct meaning. Simon and Gregg (1993) indicate that authentic assessments can "stimulate critical thought and input" (p. 6), which suggests that students are engaged in developing higher order thinking. Simon and Gregg (1993) also assert that authentic assessments "involve students in their own learning" (p. 6). These claims parallel those made for cognitive- and metacognitive-based approaches to learning.

Arguments have also been made that authentic assessment experiences encourage multiple modes of expression and support collaboration with others (California Assessment Collaborative 1993; Henderson and Karr-Kidwell 1998; Simon and Gregg 1993). Simon and Gregg also opine that authentic assessment can "increase interest" (p. 6) and "improve attitudes" (p. 6).

In summary, the increasing popularity of authentic assessment tends to parallel the displeasure with education's reliance on traditional measurement practices (e.g., standardized achievement tests). Critics argue that assessment should be more closely linked to real-world expectations and that, by reflecting the real world, resulting assessments become more valid. Therefore, validity appears to be at the heart of these claims. It should be noted that a similar concern has been addressed historically in vocational education, where standardized testing practices have been less prominent and where the boundaries between learning and assessment have been less distinct. In short, one distinct feature of vocational education is that validity concerns have been less problematic than in the more traditional academic content areas.

**Authentic Assessment—The Research**

In addition to the purported benefits of authentic assessment for the quality of student learning, some claims have also been made about the effect of authentic experiences on student interests and attitudes. Unfortunately, a review of the literature reveals a plethora of anecdotal, rather than empirical evidence. Some authors have acknowledged the rhetorical and advocacy-oriented nature of much of what has been written on the topic and have decried the lack of research. Shepard et al. (1994) state, "to date, little research has been done to evaluate the effect of performance assessments on instructional practices or on student learning" (p. 7). Concern has also been voiced about the quality of the research that has been done. This concern is illustrated in a review of portfolio research by Herman and Winters (1994). Although portfolio assessment represents only one aspect of authentic assessment, this review targeting the previous 10 years' literature on portfolios speaks volumes to the issue of quality. Herman and Winters found that, "of 89 articles written on portfolio assessment, only seven report technical data or employ accepted research methods" (p. 48). They also reported that "relatively absent is attention to technical quality, to serious indicators of impact, or to rigorous testing of assumptions" (p. 48).
Gillespie, Ford, Gillespie, and Leavell (1996) also conducted a review of the portfolio assessment literature. In that study, articles spanning the previous 5-year period were reviewed. These manuscripts had been published in the Phi Delta Kappan, Educational Leadership, and six other journals as well as two yearbooks. Although there was no attempt to distinguish between findings based on empirical research versus anecdotal reporting, the information provided was insightful. Gillespie et al. reported that “only five of the articles reviewed mentioned reliability and validity.” These results do not suggest that authentic assessment and instructional practices are invalid. Rather, although much of the rhetorical and theoretical support of authentic assessment is compelling, there remains little evidence based on empirical research to support the claims.

The Impact of Authentic Assessment on Learning

Metacognition. Metacognition is the self-management of learning by planning, implementing, and monitoring one's own learning. A metacognitive approach promoted in authentic assessment is to have students participate by using self-assessment strategies throughout the teaching-learning process. Hattie, Biggs, and Purdie (1996) conducted a study to explore this approach. A meta-analysis of 51 studies was used to determine the effect of learning skills interventions to enhance learning. Although their analysis was not limited to studies related specifically to authentic assessment, their findings support the value of metacognition. They recommend that “training for other than mnemonic performance should ... promote a high degree of learner activity and metacognitive awareness” (p. 131). This finding supports authentic assessment approaches, which call for students to participate actively in self-assessment, thereby maintaining a sense of where they have been and where they need to go.

In another study focused on metacognition and learning, Moss (1997) found that a group of elementary teachers who were exposed to a “systematic self-reflection” process (in this case, using a rubric) outperformed those who attended the same workshop but did not receive the rubric. The systematic self-reflection group tended to set goals, select interventions to match those goals, and exhibit a deeper level of understanding of the content presented. These findings have further implications for intervention practices, which require students to participate by creating assessment criteria and scoring rubrics. This suggests that allowing vocational students to participate in creating criteria for their own assessments may enhance learning.

Contextual Learning. Teaching in real-world contexts (situated learning) is another important thrust of authentic assessment. The findings of Hattie, Biggs, and Purdie's (1996) meta-analysis of learning skills interventions support the benefits of situated cognition. They recommend that training should “be in context” and “use tasks within the same domain as the target content” (p. 131). Flesher (1993) and Johnson (1987) have conducted studies on the influence of contexts when troubleshooting faults in electricity/electronics. In both studies, the results clearly support the positive influences of context on troubleshooters' abilities to locate faults.
The Value of Authentic Assessment as a Means of Assessment

**Student Self-assessment.** One of the claims that has been used to support authentic assessment is to reduce the barriers between learning and assessment. One method of doing this is to increase student involvement in their own assessment. Falchikov and Boud (1989), in a meta-analysis of 51 studies related to student self-assessment, explored the relationship between students' self-assessment (self-ratings) and their teacher's ratings. It should be noted that the studies included in their review were restricted to those providing quantitative data. The findings indicate a direct relationship between the quality of the design of the study and success of students' self-ratings. Although this illuminates the importance of designing high-quality studies, one could also infer that it is equally important to design high-quality educational activities used for authentic assessments. Another significant finding was related to the experience of the student assessors. Regarding experience and maturity, Falchikov and Boud reported that year in school (i.e., freshman, sophomore, etc.) was not found to be a significant factor in the general quality of students' self-assessments. Self-assessments of students in advanced-level courses more closely resembled their teachers' assessments than those of students in introductory courses. Therefore, when it comes to self-assessment, experience in a given field seems to be more influential than year in school.

Another interesting finding by Falchikov and Boud (1989) was that the category they termed the “broad area of sciences” produced more accurate self-assessments than did the social sciences. Although it is interesting to speculate about reasons for this, it is clear that the types of assessment experiences were relatively similar between the two groups. Also, no patterns existed to signify a difference between assessments of processes versus assessments of products. Neither were there differences between assessment of “professional practices” versus “traditional academic activities.” This last finding has direct implications for authentic assessment. “Professional practices” reflect real-world activities called for in authentic assessment. This study suggests that students do no better or worse self-evaluating these activities than they do “traditional academic activities.”

**Teachers' Level of Performance.** Another area of research has focused on how teachers are performing in the classroom. If teachers are not engaging in appropriate forms of authentic assessment, how can students be successful? Haydel, Oescher, and Banbury (1995) conducted a study designed to assess classroom teachers' performance assessments. Ninety-two performance assessments were collected from 79 teachers in a school district that was implementing outcome-based education in Louisiana. Teachers were found to have difficulty following good practices, such as defining purposes and targets and subsequently aligning the two. They also had problems articulating the performance criteria, specifying an appropriate scoring scale, and using a scoring record. It is important to note that this was a single case study, conducted in one school district. Thus, the results may not be generalizable to other populations. However, one could infer that, based on the results of this study along with the findings of Falchikov and Boud's (1989) reported in the previous section, preservice and inservice teacher
training in authentic assessment techniques and practices is likely a key factor in successful implementation.

**Reliability and Validity.** As noted in the beginning of this chapter, there is some debate in the field as to whether traditional psychometric practices (e.g., those used to establish validity and reliability) are appropriate for authentic assessment. However, given that these practices have a strong history in assessment and psychometrics, related research is examined in this chapter.

One important issue related to assessment is the ability to conclude, with confidence, that what is being reported is consistent and accurate. If policy decisions are to be made based on assessment data, it is important that the reliability and validity of the assessments be established. Gillespie et al. (1996) examined articles on portfolio assessment published over a 5-year span and found that only five mentioned reliability and validity. They thus concluded that the validity and reliability of portfolio assessment (at least for the studies examined) was “controversial at best” (p. 485).

Jiang, Smith, and Nichols (1997) conducted a meta-analysis of 22 studies published after 1980 that were found in the Educational Resources Information Center (ERIC) and Psychology Literature (PSYCHLIT) databases. The purpose of their work was to identify significant sources of measurement error influencing the reliability of performance assessment. They reported that the number one source of measurement error was due to differences in task difficulty. Further, they found that the complexity of many performance assessments often leads to multiple correct solutions. For example, in a design class, not all design problems are of equal complexity. Even when students are given the same design problem, they often come up with several different plausible solutions. Differences in tasks that have various possible levels of complexity (such as those prevalent in a design class) were found to be the most prominent source of measurement error.

The second most prominent source of measurement error was due to “occasion.” Occasion was defined as “all possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment” (ibid., p. 3). If students in a class have the freedom to choose among multiple opportunities when they are to be assessed, there will be greater opportunities for variance in grades due to measurement error. For example, if each student in a vocational welding program is allowed to choose when they are to perform a weld for a grade, there will be a greater chance for variability in grades due to measurement error.

One of the most significant findings reported by Jiang et al. (1997) was that human judgment contributed only a small amount of measurement error. They suggested that it is time that critics set aside their concerns about professional judgments involved in scoring performance assessments. Rather, their findings indicate that error due to human judgment can be minimized through training.

Another study investigated the concurrent validity of performance measures. Crehan (1997) attempted to validate a new performance measure used in a school district by investigating correlations with a norm-referenced achievement measure.
previously adopted by the district. He found no significant correlation between the performance measure and the standardized test. Although the appropriateness of using a standardized test as the validity criterion for a performance measure could be questioned, the standardized test was already accepted as a useful predictor of achievement. This is interesting considering that a major reason given for developing performance tests is the claim of the inherent limitations and weaknesses of traditional testing approaches. If one accepts this premise, then the use of standardized tests to validate performance tests could be questioned.

Parkes (1997) conducted a study that addressed the validity of a variety of testing formats with an emphasis on implications for metacognition. He attempted to determine if a student’s perceptions of control could be detected during a performance assessment. The hypothesis was that a performance assessment would provide additional information regarding student control whereas a traditional objective test on the same content would not. The findings indicated that the performance assessment score was significantly correlated to the objective test score. This finding supports the contention that they both measured similar content. The findings also indicate that the internal control scale was significantly correlated to the performance score but not significantly correlated to the objective test score. The question posed then was Did the variance due to students’ perception of internal control fall within what Messick (1992, 1996) referred to as construct irrelevant variance? Was it extra noise that needs to be controlled for during the assessment process or was it construct relevant variance that is a key part of what was trying to be measured? The researcher concluded that the objective test score measured domain knowledge whereas the performance test better measured ability to use or apply that knowledge. Because of this, Parkes (1997) concluded that “the question now is not which format is more valid, but which construct is the one we really want” (p. 10).

Summary

One of the strengths of authentic assessment is the ability to embed learning within meaningful contexts. Based on this review of the research, this contention can be supported. Teaching in context, a practice that is pervasive in vocational education, can enhance learning. This confirms what career and technology educators have known for years. What is valuable here is to have the importance of authenticity validated in areas that extend beyond vocational and technology-related areas.

The materials reviewed in this chapter also support the value of metacognitive approaches to learning and assessment. Encouraging students to become more involved in monitoring their own learning through self-evaluation can enhance student learning. From the assessment side, research indicates that students do a better job evaluating their own work in upper-level classes in a given field than entry-level classes. This could be due to a number of factors, such as maturity or additional content knowledge. Additional research is needed to explore the use of self-assessment in vocational subjects.
One significant concern throughout this review had to do with teachers' performance, both as facilitators of learning and as evaluators of student performance. Research indicates that teachers may have difficulty maintaining alignment among performance criteria, scoring scales, and the assessment records. This finding supports the need for better preservice and inservice training in authentic assessment and contextualized learning practices. It is important to note that the in-service needs of vocational teachers will likely be quite different from those of academic teachers. Contextualized learning and many authentic assessment practices are not new to vocational teachers. There is, however, an ongoing need for vocational teachers to understand how authentic assessment mechanisms work as well as how to integrate learning with academic areas.

One of the key measurement issues discussed throughout the research had to do with reliability. The two largest sources of measurement error in performance assessments were differences in difficulty of tasks and variance due to multiple occasions in which teachers are equally willing to accept a score. These represent relatively straightforward psychometric issues that must be addressed in any type of research. However, both concerns tend to be exacerbated when the emphasis shifts away from testing to context-based, authentic assessment techniques. One of the surprising findings in this review was that human judgment emerged as a less serious, and correctable, source of reliability error than might have been expected. The research indicates that proper training can minimize human judgment error. This indicates that the scoring and use of authentic assessment measures are appropriate topics for teacher inservice training.

Finally, the question of the nature of authenticity was addressed. How authentic is authentic enough? Although research indicates that context can have a positive influence on learning, there was a general lack of research investigating the ranges of authenticity. How closely does education need to mirror the real world in order to have positive impacts on learning and assessment? Is there a point of diminishing returns? Is it possible for an activity to reach a threshold of authenticity beyond which it is no longer prudent to expend the resources required to increase its effectiveness? Do all of our educational activities have to reflect the "real world"? Are there some aspects of the curriculum where learning occurs better using traditional approaches? Research remains to be done in these areas.

Authentic assessment represents an exciting attempt to stimulate learning and make it more relevant. It also represents a means for assessing students in rich and meaningful ways. Students deserve to know why it is important to learn something, and authentic teaching and evaluation methods represent a move in that direction. However, it is important to note that authentic assessments should represent only one category of tools and, like all tools, should probably not be used exclusively for all tasks.
The purpose of this chapter was to attempt to identify what the research has to say about authentic assessment. As is frequently the case in education, the linkage between practice and research is often tenuous. Trends tend to come and go. The current enthusiasm and interest in cognitive learning theory, with its emphasis on authentic learning and assessment, represents a special opportunity for vocational education. Other academic areas are coming to realize what vocational educators have known to be true for years: meaningful, contextualized experiences tend to promote better learning. The challenge remains to engage and focus the best minds in the profession to conduct the research needed to clarify how these mechanisms work…and don’t work.