Journal of Research Methodology, Volume 10, Number 1 (January - June 1997)

วารสารวิธีวิทยาการวิจัย ปีที่ 10 ฉบับที่ 1 (มกราคม-มิถุนายน 2540)

Semantics, Psychometrics, and Assessment Reform: A Close Look at "Authentic" Tests

James S. Terwilliger*

ABSTRACT

The purpose of this paper is to raise questions about the claims that are frequently made by advocates of assessment reform. In particular, the focus would be the work of Wiggins who first introduced the concept of "authentic" assessment and who was one of the most influential critics of traditional assessment approaches. The author agreed that assessment practices were in need of reform but believed that the rhetoric of the reformers was misleading and largely unsupported by data. The alternative assessment procedures should be adopted in combination with more traditional forms of assessment as evidence of the educational and psychometric value of such alternatives became aviable.

* University of Minnesota

Paper presented as part of an invited symposium on " Theory and Practice in Large - Scale Performance Assessment" at the annual of the National Council on Measurement in Education, New York City, April 9, 1996.

Introduction

It is obvious to even the most casual reader of the literature on educational assessment that the field is currently undergoing a fundamental and profoud transformation. The traditional concepts and methodologies associated with assessment are being questioned by a variety of critics including school reform advocates, subject matter experts, cognitive theorists, and others. In general, advocates for change recommend that assessments of achievement should be designed to reflect more precisely complex "real-life" performances and problems than is possible with short-answer and choice-response questions that characterize the typical standardized test currently employed in large-scale testing programs.

The purpose of this paper is to raise questions about the claims that are frequently made by advocates of assessment reform. In particular, my focus will be the work of Wiggins (1989a, 1989b, 1993) who first introduced the concept of "authentic" assessment and who is one of the most influential critics of traditional assessment approaches.

I wish to make it clear at the outset that I do not oppose some of the ideas that have been put forth by Wiggins and others. In fact, I agree that assessment practices are in need of reform. However, I believe that, as is often the case, the rhetoric of the reformers is misleading and largely unsupported by data. I fear that there is a danger that perfectly useful and appropriate assessment methods will be discarded in a rush to adopt a variety of other techniques of unknown psychometric and educational quality. I believe that alternative assessment procedures should be adopted **in combination** with more traditional forms of assessment as evidence of the educational and psychometric value of such alternatives becomes available.

Word Magic and Assessment Reform

The American Heritage Dictionary gives the following definitions of "authentic":

1. a. Worthy of trust, reliance, or belief: **authentic records.** b. Having an undisputed origin: genuine.

2. Law. Executed with due process of law: **an authentic deed.** Synonyms are listed as "real". "genuine", and "authoritative". Obviously, terms like these have a decidedly positive connotation. Therefore, any object or product to which these terms are applied is likely to be viewed as more desirable or of better quality than objects or products which are not so described.

That is why these terms appear so frequently in commercials and advertisements in the popular media, e.g. statements such as "Coke is the **real** thing!", "**Genuine** factory auto parts, and Authentic French cuisine. A bakery chain in the Twin Cities has recently introduced RENAISSANCE BREADS with the slogan "Authentic handmade breads in the European tradition." How can one doubt the quality of such a product when the price is $3.79 per loaf?

The appeal of the term "authentic" is obvious in it's widespread adoption by a variety of educators since it was first introduced into the literature. A variety of books with titles such **as Authentic Assessment in Action** (1995) and **Assessment of Authentic Performance in School Mathematics** (1992) have recently been published. Lesh and Lamon, the editors of the latter book, provide the following interesting definition:

> Stated simply, authentic mathematical activities are those that involve: (i) real mathematics, (ii) realistic situations, (iii) questions or issues that might actually occur in a real-life situation, and (iv) realistic tools and resources. (p.18)

There are two problems with this definition. First, as previously noted, words like "real", "realistic", and "real-life" are synonymous with the word "authentic". Therefore, the definition is circular. Second, and more fundamental, what constitutes "realistic" or "real-life" situations and "realistic" tools and resources is frequently open to debate. What appears to be "realistic" to one individual often seems to be "unrealistic" to another. The only way to "objectively" define "real-life" questions and situations would be to construct a data base that could be employed in defining the likelihood that a student would encounter specific questions, problems, etc. in non-school settings. Lacking such a data base, test designers typically rely on individual (or perhaps team) judgments of what is "realistic".

It seems that the word "authentic" has an almost mystical power. Phrases such as "authentic instruction" , "authentic performance ", and "authentic outcomes " are appearing with increasing frequency in the educational literature. (One can only speculate about forms of instruction, performance, and outcomes that might be labeled "inauthentic") I note that one of the NCME training sessions for this convention is titled "Ensuring that Authenic Assessment is Fair". It is reassuring that there is an increasing recognition that even "authentic" assessments pose the same problems of "fairness" as do more conventional assessment techniques.

Clearly, individuals who are not sophisticated with respect to issues associated with the design and analysis of assessment precedures are more likely to view positively any procedure which is labeled "authentic" regardless of any conceptual, technical, or practical issues that many arise in the application of the procedure. Educational assessment is a complex process which is built upon a variety of assumptions about the purposes of education along with a set of data gathering procedures which need to be judged against a series of both practical and technical standards. The use of labels which impute special status to a specific set of data collection procedures only serves to obscure more fundamental assessment questions that must be addressed. Therefore, terms like "authentic", "genuine", and "real-life" should be reserved for advertising copy and be avoided in scholarly discussions of educational assessment.

Origins of "Authentic" Tests

The term "authentic" was first introduced in reference to tests by Wiggins (1989a) in an article in **Educational Leadership**, a journal for school administrators and general educators. (This is an audience unlikely to include many psychometrically trained readers.) Wiggins defined "authentic" tests in terms of complex performances or exhibitions in which a student completes a report or makes a public presentation following an extended period of work on an out-of-class assignment. Wiggins presents the example shown in Figure 1.

Fig.1. An Example of a Test of Performance

An Oral History Project for 9 th Graders

To the student:

You must complete an oral history based on interviews and written sources and then present your findings orally in class. The choice of subject matter is up to you. Some examples of possible topics include: your family, running a small business, substance abuse, a labor union, teenage parents, and recent immigrants.

Create three workable hypotheses based on your preliminary investigations and four questions you will ask to test out each hypothesis

Criteria for Evaluation of Oral History Project

To the teacher:

Did student investigate three hypotheses?
Did student describe at least one change over time?
Did student demonstrate that he or she had done background research?
Were the four people selected for the interviews appropriate sources?
Did student prepare at least four questions in advance, related to each hypothesis?
Were those questions leading or biased?
Were follow-up questions asked where possible, based on answers?
Did student not important differences between "fact" and "opinion" in answers?
Did student use evidence to prove the ultimate best hypothesis?
Did student exhibit organization in writing and presentation to class?

Note: This example is courtesy of Albin Moser, Hope High School, Providence, Rhode Island. To obtain a thorough account of a performance-based history course, including the lessons used and pitfalls encountered, write to Dave Kobrin, Brown University, Education Department, Providence, RI 02912.

From Wiggins, 1989 (p. 44)

Wiggins concludes his discussion of "authentic" tests as follows:

> In sum, authentic tests have four basic characteristics in common. First, they are designed to be truly representative of performance in the field; only then are the problems of scoring reliability and logistics of testing considered. Second, far greater attention is paid to the teaching and learning of the criteria to be used in the assessment. Third, self-assessment plays a much greater role than in conventional testing. And fourth, the students are often expected to present their work and defend themselves publicly and orally to ensure that their apparent mastery is genuine. (p.45)

It is instructive to examine the example Wiggins gives in light of the four characteristics he claims all "authentic" tests share. First, what " field " is represented in the oral history project? Since the choice of topic is left to the student, the "field" must be history with special emphasis upon techiques employed by historians who employ "first-person" sources in their research. In fact, this approach is hardly "truly representative of performance in th e field" if the "field" is more broadly defined as history since the great majority of historians rely upon written rather than "first-person" source in their work.

Second, it is not clear that the criteria for evaluation of the project were shared with the students.(In fact, the example strongly suggests that the criteria were for the teacher only.) Therefore, how could they have been "taught and learned"? Even if the cirteria were shared with the students in advance, what exactly would a teacher expect them to learn from them? Most students would use the criteria as a checklist to make certain they had satisfied the teacher's demands prior to turning their projects in for evaluation by the teacher! Since many of these criteria are very specific to the particular project, they appear to have limited value as general learning outcomes.

As an aside, it is not clear from the criteria presented exactly how they are to be employed in evaluating the projects. Since most of the questions posed can readily be answered "yes" or "no", it would be possible to devise a simple checklist. Obviously, a more elaborate scoring system could also be designed but the example provides no clues if that is the case. As is the case in all examples he gieves in his writings, Wiggins provides no data with respect to the reliability of scoring (or the amount of time devoted to scoring) the "authentic" tasks he recommends. The lack of supporting data is a reflection of the secondary role Wiggins gives to such issues in his reference to "scoring reliability and logistics of testing" in the above quote.

Third, with respect to self-assessment, there is no indication that this is involved in the oral history project.

Lastly, it is clear that the oral history project does involve an oral presentation to the class. Presumably, the oral presentation may also be followed by a discussion during which the presenter would have to answer questions and defend her/his work.

"Authenticity" and Validity

Several issues are highlighted through the detailed comparison of the oral history example with the criteria for "authentic" tests given by Wiggins. (The same example and criteria for "authenticity" are presented in a follow-up article (1989b) which appeared in the **Phi Delta Kappan,** another journal for a general readership in education.) First, it is not entirely clear what is meant by the phrase, "designed to be truly representative of performance in the field". What exactly is the "field" in this example? Who decides what is "truly representative" of the field?. This raises questions regarding test **validity.** In his follow-up article Wiggins' only reference to validity of "authentic" tests is the comment, "Far greater attention is paid throughout to the test's 'face' and 'ecological' validity." (1989b,p.712) However, in his more recent writings, Wiggins acknowledges that "face" validity is not highly regarded by measurement specialists and states,

> Thus, although face validity should be considered, to focus only on it is to miss a more important point about the incentives to perform well that might be found to inhere in more authentic forms of assessment and might change the implications of the scores. (1993, p.244)

This statement seems to imply that interpretations (hence the validity) of results of "authentic" assessments is somehow enhanced by virtue of their being perceived as more "genuine" than conventional forms of assessment. Of course, this is one of the major arguments frequently advanced in support of "face" validity.

Messick (1994) has discussed at length issues associated with the validation of performance assessments. He cites the classic treatment of performance tests by Fitzpatrick and Morrison (1971) in which they state, "there is no absolute distinction between performance tests and other classes of tests" (p.238) Using this as a point of departure, Messick states,

> Hence performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as are other assessments. Indeed, such basic assessment issues as validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are **social values** that have meaning and force outside of measurement wherever evaluative judgments and decisions are made. (p.13)

With regard to special claims of "authenticity" , Messick notes,

> The portrayal of performance assessmemts as **authentic** and **direct** has all the earmarks of a validity claim with little or no evidential grounding. That is , if authenticity is important to consider when evaluating the consequences of assessment for student achievement, it constitutes a tacit validity standard, as does the closely related concept of directness of assessment. We need to address what the labels authentic and direct might mean in validity terms. We also need to determine what kinds of evidence might legitimize both their use as validity standards and their nefarious implication that other forms of assessment are not only indirect, but inauthentic. (p.14)

As previously noted, Wiggins presents **no** validity data, evidential or consequential, in any of his writings on "authentic" assessment.

Assessment and Educational Philosophy

A theme that runs through Wiggins' conception of "authentic" testing is an emphasis upon performances that are designed to assess "higher order" outcomes. Among the design features he lists (1989a,p.45) as characteristic of "authentic" tests are: contextualized, complex intellectual challenges; involves student's own research; emphasizes **depth** more than breath; and involves somewhat "ill-structured" tasks or problems. He specifically disavows any interest in "atomized" tasks, corresponding to isolated "outcomes", "mere recall" , and "plug-in skills". The reason for this is made clear in his discussion of validity in his 1993 book.

> At bottom is a philosophical problem of major proportions about the purpose of schooling: Is schooling meant to yield common knowledge? If so, then it makes perfect sense to think of tests as properly focusing on what students hold in common . But what if education is a personal, idiosyncratic affair, where the meaning and personal effectiveness that I derive from coursework is more important than what knowledge we all end up holding in common? In that case, a standardized, indirect test- -of any kind- -would make no sense: What could we possibly mean by a standardized test of meaning of educative experience? What, then, of the validity of aptitude tests if I cannot state with precision: aptitude **for what future role or aspiration?** (p. 247)

Wiggins (1993) gives examples of a variety of "roles and situations through which students can perform with knowledge". The roles include such diverse occupations and settings as follow (p.223-24):

museum curator
U.N representative
tour organizer
bank manager
historian
engineering designer
ad agency director
psychologist/sociologist
newspaper editor and writer
product designer

job applicant
expert witness
reviewer
teacher
speaker-listener
commercial designer

He argues that such roles could serve as "template" for better test design because, "These roles and situational challenges are common to professional life." Wiggins claims it is a "logical fallacy" to argue that students must be given "drills" and " tests concerning their mastery of the drills" prior to requiring performance in professional roles and situations . "Drill testing" is a means to an end and "it is certainly not to be confused with the important performance itself".

Wiggins conveniently ignores the possibility that most 'roles and situational challenges common to professional life' involve an extensive knowledge base. Individuals who lack the knowledge base have little or no chance of performing successfully in the' real-life' roles which he describes. For example , an historian typically specializes in a particular time period and geographical region for her/his research, e.g. Colonial America during the period 1650-1770. In order to make a useful contribution to the literature on this topic, the historian must first become familiar with a vast amount of work previously published by other historians working on this and related topics.

The Importance of Knowledge

Knowledge and the Taxonomy

I believe that phrases like "drills" and "drill testing" are employed with the intent of denigrating the role of **knowledge** in the various "roles" and "situational challenges" to which Wiggins refers. In fact, knowledge is a basic building block upon which most "higher order" educational outcomes rest. This is explicitly acknowledged in the most widely cited framework for classifying educational outcomes, the Bloom taxonomy, which defines"knowledge" as the most elementary of all the cognitive outcomes.

It is interesting to note that Wiggins claims that the authors of the taxonomy did not intend to imply that educational outcomes necessarily occur in a "fixed chronological sequence" beginning with knowledge. This claim is made despite the fact that the taxonomy is clearly presented by its authors as an **hierarchical** model wiht outcomes arranged on a simple- to-complex scale of cognitive processes.

It is also important to recall that "knowledge" as employed by the authors of the taxonomy extends well beyond conventional notions of "facts". Subcategories in the taxonomy include the following:

Knowledge of terms.
Knowledge of facts.
Knowledge of concepts.
Knowledge of principles.
Knowledge of procedure.
Knowledge of theories.

This list of diverse outcomes should make it clear that the "knowledge" intended by the authors of the taxonomy is not necessarily limited to the type of "knowledge" achieved strictly through "drills" and "drill testing"

Knowledge and Expertise

The fundamental role played by knowledge in "real-life" roles is well documented in the extensive body of work on the nature of expertise. Chi, Glaser, and Farr (1988) have edited a series of papers which summarize much of the work in this field. In their overview of this work Glaser and Chi list several "key characteristics" of the performance of experts. The **first** characterstic they cite is that expertise generally is restricted to specific domains of performance and does not transfer from one domain to another. They state,

> The obvious reason for the excellence of experts is that they have a good deal of domain knowledge. This is easily demonstrated; for example, in medical diagnosis, expert physicians have more differentiations of common diseases into disease variants (Johnson et al., 1981). Likewise, in examining taxi drivers' knowledge of routes, Chase (1983) found that expert drivers can generate a far greater number of secondary routes (i.e. lesser known streets) than novice drivers. (1988, p. xvii)

Knowledge and Literacy

Wiggin's emphasis upon roles "common to professional life" reflects a very narrow view of educational outcomes. Surely, there are many other outcomes of education which are also important, e.g. actively participating in the economic, political, and cultural life of one's community; acting as a responsible member of a family and various other social organizations; making contributions to the improvement of the life of one's community through involvement in civic and volunteer organizations, etc. In other words, it is also important that students develop into adults who take their roles as citizens seriously and who make positive contributions to the improvement of society.

Hirsch (1987) argues that a fundamental goal of education shoud be to produce students who are "culturally literate". His bestselling book, **Cultural Literacy-What Every American Needs to Know,** was written in response to several studies which indicated a serious decline in the level of knowledged possessed by American students durintg the 1970's and early 1980's. Hirsch describes cultural literacy as the "network of information that all competent readers possess".

> It is the background information stored in their minds, that enables them to take up a newspaper and read it with an adequate level of comprehension, getting the point, grasping the implications, relating what they read to the unstated context which alone gives meaning to what they read. (1987,p.2)

Hirsch is critical of attempts to dismiss knowledge in favor of more lofty educational goals. He argues that the denigration of "mere facts" by advocates of instruction in "higher order" skills creates a false dichotomy.

> The polarization of educationists into facts-people versus skills-people has no basis in reason. Facts and skills are inseparable. There is no insurmountable reason why those who advocate the teaching of higher order skills and those who advocate the teaching of common traditional content should not join forces. No philosophical or practical barrier prevents them from doing so, and all who consider mature literacy to be a paramount aim of education will wish them to do so. (1987, p.133)

Proponents of educational reform who stress "critical thinking" and similar "higher order" thinking skills should consider Hirsch's advice seriously when planning classroom instruction and assessment.

The Price of Ignorance: A Recent Warning

(This section is based largely upon an article by Richard Morin of the **Washington Post** which appeared in the **Minneapolis Star Tribune** on February 12, 1996.)

There are many studies which document the lack of knowledge that characterizes American students on topics ranging from math and science to history and geography. A recent survey conducted by the **Washington Post,** the Kaiser Family Foundation, and Harvard University reveals a similar lack of knowledge in the area of U.S. government and politics. The study authors interviewed 1,514 randomly selected adults (not students) in November and December, 1995. Study participants were asked 18 general knowledge questions about the government and leaders in government. An additional 21 questions concerning political knowledge were asked in four other national surveys conducted by the **Post**. Some selected findings are as follow:

> two-thirds of those interviewed could not name their representative in the U.S House of Representatives
>
> 40% did not know the name of the Vice President
>
> two-thirds could not name the majority leader of the U.S. Senate
>
> 46% did not know the name of the speaker of the House of Representatives
>
> 75% were not aware that U.S. senators are elected for six- year terms

40% did not know that Republicans control both chambers of Congress

nearly 60% incorrectly believed that the government spends more on foreign aid than Medicare

The survey findings further suggests some of the more important consequences of lack of knowledge. Those who are less informed are much more likely to tune out politics and turn off voting. Also, those who are less informed are more likely to believe that the country is in decline. For example, less informed Americans are far more likely to believe that that annual budget deficit and the number of federal employees had increased- - not decreased--in recent years.

The gap in information also affects how politics is practiced according to Samuel Popkin, a political scientist at U.C., San Diego. He states that candidates for office run two distinct campaigns; one for informed voters which stress their stands on issues and policies and one in which strategists and consultants aim attacks at the character of the opponent to win support from less-informed voters. Popkin claims that , for the less- informed voter, all politics is reduced to character and caricature: politicians are separated into "heroes and villains"and major policy debates become clashes between good and evil. The existence of a substantial block of poorly informed voters is one reason why so-called 'negative campaigning' seems to be so effective

A Case History of an "Authentic" Test

The notion that complex measures of achievement be used to assess performance in more a "lifelike" setting than is represented by conventional standardized tests is not new. In fact, instruments designed to be more "authentic" than paper-and - pencil exams have been employed in particular occupational and educational setting for many years .[1] The in-basket test is the product of over five decades of research and application in business and managerial settings. It is a simulation exercise using materials typically found in a hypothetical managerial position. The materials in the in-basket consist of letters, memos, records, and other items which require a response from the examinee who role-plays as the occupant of the hypothetical position. The in-basket methodology is frequently used as an assessment tool for predicting job performance.

[1] Swanson, Norman, and Linn (1995) recently reviewed the long history of such measures in the health prefessions. Their article describes four specific performance-based methods commonly employed in training health-care professionals and lists "some hard lessons learned from many studies and frequent missteps". Advocates of the use of complex performance measures in large scale assessment programs would be well advised to study the eight"lessons" listed in this article.

Schippmann, Prien, and Katz (1990) did a comprehensive review of the literature on the psychometric properties of in - basket tests. They examined all available studies that reported reliability and/or validity data using in- basket tests. Their findings are as follow:

> Interrater reliability-- "it appears that scorers and raters are responding fairly consistently to participant response data. The range of interrater values, however, suggests that something else is operating which creates deviant score/rating patterns. This'something else' may well be a function of scorer/rater training."
>
> Alternate-form reliability--"The difference between obtaine d alternate-form reliabilities and interrater reliabilities is dramatic. In the absence of additional research data, the conservative conclusion is that the consistency of individual performance is, at best, marginal."
>
> Split-half reliability--"split-half reliabilities of in- basket performance measures have been somewhat disappointing. Though it would appear that greater rigor in the development of test content and more systematic and objective scoring procedures might yield more encouraging reliabilities,additional research is needed in order to draw firm conclusions."
>
> Content validity--"all of the reviewed studies which suggest that their procedures are content valid fall seriously short of the mark in our view."--"There simply are no published or widely distributed reports which describe how to develop an in- basket that is well grounded with regard to content validity."
>
> Criterion-related validity--"studies of criterion-related validity did reveal a large number of significant correlations between in-basket measures and various criteria. Thus the evidence of criterion-related validity for certain in-baskets is sufficient to support the development and use of the procedure for various decision making purposes."--"However, it is frequently difficult to determine the in-basket's contribution to the prediction of performance apart from other assessment tools."
>
> Construct validity--"evidence of construct validity was suggested in those cases where the in-basket was designed to reflect some theoretical construct. The conclusion warranted here is that the evidence from these studies is encouraging, though not convincing with reference to either hypothetical constructs or job performance criterion constructs."

Schippmann, et al. (1990) make some observations in their discussion of validity that have special relevance for advocates of "authentic" tests.

> the issue of the content validity of in-basket exercises appears to be clouded by confusion between face validity and content validity. Simply eyeballing test content with reference to the job analysis results and proclaiming the procedures to be content valid is unacceptable. The appearance of the relevance may contribute to face validity, though it is not a sufficient basis for making inferences about job performance . (p.851)

Conclusions

Wiggin's promotion of "authentic" assessment, however well intentioned, is flawed in several important respects. First , the term "authentic" is misleading and confusing. The term inappropriately implies that some assessment approaches are superior to others because they measure outcomes that are more "genuine" or "real". This claim is based largely upon an appeal to "face" validity , a concept which has been abandoned by modern psychometric theorists. Second, Wiggins rejects the role of knowledge in the assessment of educational outcomes. This ignores a substantial body of theory and ample empirical evidence which supports the central role of knowledge in countless domains of performance. Third, the presumption that complex performance-based measures of assessment can be designed to achieve the psychometric quality necessary for use in large-scale assessment programs is , at best, questionable in light of previous efforts to employ such measures in other professions.

References

Darling-Hammond, L.,Ancess,J., and Falk, B. (1995) ***Authentic Assessment in Action: Studies of Schools and Students at Work.*** New York: Teacher College Press.

Chi, M., Glaser, R., and Farr, M. (Eds.) (1988). ***The Nature of Expertise.*** Hillsdale, NJ: Lawrence Erlbaum Associates.

Fitzpatrick, R. and Morrison, E. (1971). Performance and Product Evaluation. In R.L. Thorndike(Ed.) ***Educational Measurement,*** 2 nd Ed. New York: American Council on Education/Macmillan.

Hirsch, E.D. (1987). ***Cultural Literacy: What Every American Needs To Know.*** New York: Houghton-Miffin Co.

Lesh, R. and Lamon, S. (Eds.) (1992) ***Assessment of Authentic Performance in School Mathematics.*** Washington: AAAS Press.

Messick,S.(1994). The interplay of evidence and consequences in the validation of performance assessment. ***Educational Researcher, 23*** (2), 13-23.

Morin,R. (1996). Of politics, government: Citizen ignorance rampant, costly. ***Minneapolis Star Tribune,*** pg. A5, Feb. 12.

Schippman, J., Prien, E., and Katz, J. (1990). Reliability and validity of in-basket performance measures. ***Personnel Psychology, 43*** (4), 837-859.

Swanson, D., Norman, G., and Linn, R.(1995) Performance-based assessment: Lesson from the health professions. ***Educational Researcher, 24*** (5), 5-11,35.

Wiggins, G.(1998a). A true test: Toward more authentic and equitable assessment. ***Phi Delta Kappan, 20,*** 703-713.

Wiggins, G. (1989). Teaching to the (authentic) test. ***Educational Leadership, 46.*** 41-47.

Wiggins, G. (1993). ***Assessing Student Performance.*** San Francisco: Jossey- Bass Publishers.