Authentic Language Testing Design: A Practical Approach

Greg Brakefield

Abstract

Authenticity in language testing has been a subject of much debate and to date, the issue still remains more theoretical than practical. In this paper I posit that it is in fact possible to design a language test that is largely authentic. I will look at the rudimentary facets of test design and will apply a conceptual framework to each facet in order to give language professionals a basic template from which to begin the process of designing an authentic language test.
Keywords: Authenticity, Language Test, Test Design, Construct, Operationalize

Introduction - What is Authenticity?

When addressing the statement, "*It is impossible to design a truly authentic language test,*" a variety of issues must first be dealt with. The statement itself begs the question, which is; what does the word "authentic" mean in the larger context of language testing? Likewise, in the smaller context, what does the word "truly," mean? In both cases, these words are constructs and as such, need to be defined so that they can be objectively measured to test for validity and reliability in order to know if the test is actually testing what it is supposed to and how well it is doing it.

When discussing authenticity in language testing Carroll (1961) was one of the first proponents of testing which required an integrated and facile performance on the part of the test taker, which focused on the total communicative effect of an utterance as it would in a non-test situation. This could be considered one of the earliest prescriptions for authenticity and a starting point for the operationalization of authenticity.

Later, Bachman (1990) defined authenticity in terms of the 'Real Life' or RL

approach, which considers the extent to which test performance replicates some specified non-test language performance or mirrors the "reality" of non-test language use in a test situation for the purpose of determining predictive utility. Bachman also defines authenticity in terms of the 'interactional/ability' or IA approach, which focuses on the interaction between the language user, the context, and the discourse as opposed to non-test language performance in the RL approach. Clark (1972) describes the criteria for authenticity in terms of proficiency or the ability to do X in reference to real life.

In a later paper, Bachman & Palmer (1996) define authenticity in terms of situational authenticity i.e. the perceived match between the characteristics of test tasks to target language use (TLU) tasks or in terms of interactionally authentic i.e. the interaction between the test taker and the test task. Finally, Bachman himself concedes that (1990 p330) "The characterization of authenticity is undoubtedly one of the most difficult problems for language testing, as it necessarily involves the consideration of not only the context in which testing takes place, but also the qualities of the test taker and the very nature of language ability itself."

Over time, the operationalization of authenticity has become more precise, but it still lacks the minimum critical mass to be considered a well operationalized concept. This is largely due to the fact that regardless of which definition/ operationalization or domain one chooses to use, the common denominator is "real life" which can be seen as a metaphor for the infinitely complex domain of human interaction/communication and therein lies the problem; how does one operationalize something as complex as the infinite variety human interaction/ communication to a sufficient degree so as to achieve a degree of validity and reliability in testing?

Which leads next to the operationalization of the word, "truly." Dictionaries provide various meanings ranging from *sincerely, frankly to certainly* and *beyond a doubt.* In the context of language teaching, meanings such as *certainly* and *beyond a doubt*, which are absolutes, would be the likely choice. There are few absolutes in life and fewer still in regards to what could be considered an authentic language test. That being the case, it is unlikely that a "truly" (a test that is absolutely authentic) authentic language test can be designed. However, this does not mean that authentic

language tests, which are useful, though not absolutely/truly authentic, could not be designed.

Construct Validity

When discussing authenticity and its importance with regards to language test design and its relationship to construct validity, we must first define it. Construct validity, as Hughes (1989) notes, refers to the extent to which a test measures a theoretical construct or trait. Communicative language tests rely on or attempt to rely upon the construct of authenticity in order to measure either directly or indirectly some aspect of language communication. This theoretically derived notion in turn drives content validity, which refers to the extent to which the content of a test covers a representative sample of the behavior to be measured. To have high content validity Weir (1990) argues that a test should contain tasks and text types which are similar to those which candidates would have to undertake in their future domain of language use i.e., which tap the type of proficiency described in the test specifications. It is obvious then that content and construct validity are closely related in as much as the theory (construct validity) drives the design of tests which are a representative sampling of real life or content validity.

Ultimately, the relationships between authenticity, construct validity and content validity are important in language test design because as Alderson (1981) states, "However one evaluates any theory, presumably by its operationalization, if operational definitions are not possible, then the theory is poorly stated or inadequate." That being the case, the idea of authenticity as a well operationalized construct is crucial to achieve sufficient content validity which in turn helps ensure that language tests are accurate in measuring the communicative language abilities that are sought after.

Having conceded that is not possible to design a "truly" authentic language test, it is nevertheless possible to design a language test that approximates authenticity and is thus useful. However, designing such a test is still a challenging proposition, and issues such as validity, reliability, and practicality must be addressed.

Direct vs. Indirect Testing

In designing an authentic language test, first, the type of test must be determined. Tests that use authenticity as a construct are more likely to be communicative in nature, focusing on speaking. In that case, a direct test that measures proficiency, (in which reasonably wide sampling is done), would be preferable to an indirect test, which measures the abilities that underlie the skill Hughes (1989). A direct test would also be one which attempts to duplicate as closely as possible the setting and operation of the real-life situations in which the proficiency is demonstrated Clark (1975). The key phrase in Clark's idea is the relative idea of "duplicating as closely as possible, the real-life setting".

Practical issues of resources such as money and time will limit the degree of replication of a real-life situation in the test. Furthermore, issues such as defining the complex nature of 'real-life' language use as well as the contexts it occurs in will ultimately place limitations on test design that will limit how authentic the test is Bachman (1990). In addition Skehan (1984:208) observes that issues of sampling can also contribute to the previously mentioned limitations, noting that because "an interaction is 'authentic' does not guarantee that the sampling of language involved will be sufficient or the basis for wide ranging and power predictions of language behaviour in other situations".

Despite these limitations, direct testing at this point in time relative to current theory is still a preferable choice to indirect testing for the simple fact that it could measure (under carefully designed, limited circumstances) proficiency better than indirect testing.

Discrete Point vs. Integrative Testing

The next step in designing an authentic language test would be to choose between discrete point and integrative testing. Hughes (1989) notes that in direct testing, the integrative approach is preferable in that it requires the candidate to combine many different language elements in the completion of tasks i.e., answering an interview question, which would be a good way to demonstrate oral proficiency. Integrative testing, because of its more complex nature is better suited for designing

authentic tests, than discrete point testing, which focuses on individual test elements which are very narrowly defined such as a multiple choice grammar test.

Norm Referenced vs. Criterion Referenced Testing

Choosing between norm referenced testing and criterion referenced testing would be the next step in designing an authentic language test. Criterion referenced testing would be preferable as it measures individual ability and or satisfactory performance as opposed to performance in comparison to group norms. However, there are inherent difficulties when using CR tests in testing communicative language ability, which are largely related to difficulties in specifying domains with respect to the real-life approach to authenticity, specifically, "…identifying the essential characteristics of such tests and defining these characteristics in a way that is consistent with considerations that must be made with respect to validity and authenticity." Bachman (1990). While this is complex, Bachman further notes that in institutional settings where domains can reasonably be specified, CR tests are particularly relevant i.e., achievement testing. Despite inherent complexities given the current state of theory and test design, CR testing is still preferable to norm referenced testing when designing an authentic language test.

Subjective vs. Objective Testing

Lastly, a choice between subjective and objective testing must be made. By its very definition, authentic language tests, which approximate real life communication, will need to be judged on a largely subjective (which is limited as much as possible) basis. Hughes (1989) gives a number of criteria to be used for the purpose ameliorating inconsistencies caused by subjectivity:

1) Take enough samples of behaviour
2) Exclude items which do not discriminate well between weaker and stronger students
3) Do not allow candidates to much freedom
4) Design unambiguous items
5) Provide clear and explicit instructions
6) Ensure that tests are well laid out and legible

7) Make candidates familiar with format and testing techniques
8) Provide uniform and non-distraction conditions of administration
9) Use items that permit scoring which is objective as possible
10) Provide a detailed scoring key
11) Train scorers
12) Agree on acceptable responses and appropriate scores at the outset of scoring
13) Identify candidates by number, not by name
14) Employ multiple independent scoring

Some of these criteria will have limitations in designing an authentic communicative language test, notably items 3, 9, 10 & 12. Item number 3 is counter-intuitive if one wishes to approximate a real life context. Items 9, 10 & 12 are enormously complex tasks and as such will be limited in the degree with which they are successful. Regardless though, these prescriptions as a basic starting point are sound in that they all work together to varying degrees to increase objectivity and adequately balance the tension between reliability and validity Hughes (1989).

Test Validity, Reliability & Practicality

The previous section covered the various criteria for designing an authentic language test such as choosing between direct and indirect testing, discrete point and integrative, norm referenced and criterion referenced, subjective and objective as well as a prescriptive list of guidelines to follow in the design of an authentic language test. However, all of these components must be considered and assembled against the larger backdrops of validity, reliability and practicality.

In designing an authentic language test, a three-dimensional approach to validity needs to be taken. Numerous writers such as Davies (1968b) ; Spolsky (1968) ; say that the prediction of future performance is the primary purpose of proficiency tests. Wesch (1985) notes that the predictive utility (a function of validity) is expected to predict how successfully the examinee will be able to communicate using the second language in certain target situations, which can be construed as an authentic language test. Therefore, this aspect of validity is quite important when considering how to design an authentic language test.

However, Clark (1978b) notes that content validity is the most important concern with direct tests in that, "The formal correspondence between the setting and the operation of the testing procedure and the setting and operation of the real-life situation constitutes the face/content validity of the test - the basic psychometric touchstone for direct proficiency tests". While content validity is of great importance, it is quite difficult to obtain an adequate sample from the relevant domain to be tested Bachman (1990). Clark (1978b) further notes, "The specification of test content, in virtually every instance must involve sampling from an extremely large number of potentially testable elements. Unfortunately, the identification of meaningful domains is an extremely complex matter."

In addition to face validity/predictive utility and content validity, construct validity is also quite important because as Bachman (1990) notes, "predictive utility is essentially precluded without authenticity." Therefore, test designers must delicately balance face validity, content validity and construct validity to varying degrees in order to design an authentic language test but as previously mentioned, this is difficult to achieve because of the complexities surrounding adequate construct and domain definition as well as adequate sampling.

While validity would generally be considered to take precedence in importance to reliability, reliability, along with validity, is still one of the essential measurement qualities (Bachman & Palmer 1996). Bachman further notes that completely eliminating inconsistencies is not possible; these inconsistencies can be minimized through test design. Therefore, by focusing on adequate test design and by taking care with regards to achieving adequate validity, reliability it could be assumed, will follow.

Validity and reliability are two essential measurement qualities, though practicality often takes precedence over them Bachman (1990). Bachman & Palmer (1996) define practicality as available resources divided by required resources. These resources are composed of human resources, material resources and time resources, and are all balanced against the backdrop of test design which attempts to achieve adequate authenticity, validity and reliability; a difficult proposition. The constraint of practicality is particularly notable when designing authentic language

tests in that these tests attempt to replicate or mimic the complexities of real-life communication.

Attempts to replicate this would ideally involve extended observation (or something similar) which is an interesting notion but as many writers recognize, it is time consuming, cumbersome, expensive and raises ethical questions as well and is therefore impractical. In this sense, practicality, perhaps unlike validity and reliability, while difficult to manage and achieve, is something less of a conundrum and represents a different kind of difficulty related to the pragmatic nature of resource management, which lies somewhat outside the more theoretical realm of authentic language test design.

Usefulness & Practicality

Well aware of the significant challenges of designing an authentic language test which focuses on maximizing the sometimes contradictory nature of the various components of reliability, construct validity, authenticity, and practicality Bachman & Palmer (1996) have come up with a novel way which seeks to find an idealized balance (usefulness) between these individual items as a group rather than focusing on the maximization of each individually. This notion is guided by three principles:

1) It is the overall usefulness of the test that is to be maximized, not the individual elements.
2) The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.
3) Test usefulness and appropriate balance among the different test qualities cannot be prescribed in general but must be determined for each specific testing situation.

This is an interesting solution to addressing the inherent complexities of designing an authentic language test in that it concedes the inherent difficulties associated with achieving adequate construct validity, reliability etc. and focuses more on a manageable, real world approach. However, Bachman & Palmer (1996) do concede the point that evaluation of a given test is essentially subjective on the part

of test developers and that depending on the individual test situation such as large groups or small groups, high stakes or low stakes, test designers will need to choose between focusing more on some components which in turn means focusing less on others, resulting in a somewhat less balanced result than Bachman & Palmer had likely hoped for.

Summary

It can be seen that designing a 'truly authentic' language test is not a likelihood at present due to the difficulties in managing the complexities of defining constructs and domains while simultaneously trying to achieve adequate validity, reliability, and sampling while balancing practical concerns of resource management. But that is not to say that language tests cannot be designed which are nevertheless close approximations to authenticity and are thus a useful and necessary evolution in current test design.

References

Alderson, J. C. (1981). 'Report on the discussion on communicative language testing'. In Alderson and A. Hughes (eds.). 'Issues in language testing'. ELT Documents 111. London: The British Council.

Bachman, L. F. (1990). 'Fundamental considerations in language testing'. Oxford: Oxford University Press.

Bachman, L. F. and A. S. Palmer (1996). 'Language testing in practice'. Oxford: Oxford University Press.

Carroll, J. B. (1961a). Fundamental considerations in testing for English proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30-40). Washington, D.C.: Center for Applied Linguistics

Clark, J. L. D. (1972). *Foreign Language Testing: Theory and Practice*. Philadelphia, PA,: Center for Curriculum Development, Inc.

Clark, J. L. D. (1975). 'Theoretical and technical considerations in oral proficiency testing' in Jones and Spolsky 1975: 10-24.

Clark, J. L. D. (1978b). 'Interview testing research at Educational Service' in Clark 1978a: 211-28.

Davies, A. (1968b). *Language Testing Symposium. A Psycholinguistic Perspective.*

London: Oxford University Press.

Hughes, A. (1989). 'Testing for Language teachers'. Cambridge: Cambridge University Press.

Lewkowicz, J. (2000). Authenticity in Language Testing. *Language Testing* 17. 1:43-64.

Skehan, P. (1984). 'Issues in the testing of English for specific purposes.' *Language Testing* 1, 2: 202-20.

Spolsky, B. (1968). 'Language testing: the problem of validation.' *TESOL QUARTERLY* 2, 2: 88-94.

Weir, C. J. (1990). 'Communicative language testing'. London: Prentice Hall.

Wesch, M. (1985). 'Introduction' in Hauptman *et al.* 1985: 1-12.