Background Research Paper No. 13

Content Analysis of *Fountas & Pinnell Benchmark Reading Assessment* When Used Cross-Culturally

Paul L. Landry  
The University of Alabama

The purpose of this content analysis is to examine the *Fountas & Pinnell Benchmark Reading Assessment for Independent Reading* Levels in which reference language is used to evaluate the language and reading comprehension level of Kindergarten-8 students. This assessment tool is part of a system for evaluating student reading level, and other parts of the system are used for differentiated instruction based upon the benchmark level designation of the student. The concept is to match student reading material for literacy instruction to the most appropriate level of difficulty for the student’s development on what the authors term a “continuum of literacy learning.”

**Purpose of Content Analysis**

The intent of this content analysis is not to critique the structure or purpose of the Fountas & Pinnell system, per se. Rather, concerns explored here relate more to commercialization and mass marketing of a system designed for a specific purpose and in a specific context. Generalized usage across cultures and for populations in different countries where backgrounds and cultural meanings may differ substantially from foundations upon which the system was constructed may be inappropriate. Without careful consideration of the context in which the system is to be applied, assessments and derived inferences may be biased and damaging to the students assessed. More broadly, available teacher materials in reading, and systems, may suffer from lack of attention to the diverse characteristics of students now expected to attain higher levels of reading comprehension.

Application of the Fountas & Pinnell system is predicated upon determining the student’s level of language and reading comprehension. Thus, the relative accuracy of the assessment is critical to the most effective use of the system. The underlying premise is that material too difficult or too easy for the student, *i.e.* a mismatch to his or her reading comprehension level, will compromise instruction by lowering student engagement. Consequently, if the assessment tool fails to properly identify the student’s comprehension level because of bias in content or interpretation of the assessment, application of the system will be less effective. The system has received broad acceptance as an effective strategy for reading recovery and literacy instruction in the U.S. public schools in which it has been tried.
The system also is being applied in international school settings where the native language of students is not English, but it is the language of assessment administration. Where references in the test items may be culture bound or unfamiliar to the culture of the students being assessed, the standard calibration of assessment would disadvantage the student, and potentially place the student at a comprehension level not accurately reflecting the student’s general understanding of the material. A related purpose of this content analysis, therefore, is to consider whether the Fountas & Pinnell system, which may function well in the context of the U.S. educational system, is inappropriate for use in multicultural or international contexts involving English Language Learners (ELL).

Literature Review

The research and review of literature will have three main areas of focus: (a) proposed methodology for evaluation of test item bias, (b) test bias with respect to language and cultural differences when testing for reading and language comprehension, and (c) limitations of effective cross-national and cross-cultural inferences to be drawn from the use of standardized assessment results. For purposes of the content analysis proposed here, the differential item function (DIF) model seems more appropriate. A brief overview of research on development and use of DIF models to evaluate assessments is provided. Cross cultural bias has received increasing attention and research effort. With respect to the last of these areas, cross-national usage and comparative inferences from test results, there is less published research available. As commercialization of standardized testing increases and cross-national use is promoted for economic gain, however, greater attention to the implications of such use is warranted and further research would be helpful.

Analytical Framework: Accuracy of Measurement and Item Bias

Differential Item Function (DIF) is a method of analysis used to identify test items that have different levels of difficulty or to which correct responses are less accessible to test participants of the same knowledge, skill, or ability level based upon factors that should be irrelevant to test performance (Ferne & Rupp, 2007). Researchers debate the use of the term “differential item function” instead of the shorter and more convenient term, “bias.” However, differential functioning is a prerequisite for bias, but insufficient to establish bias. An item that does not show DIF is not biased, but if DIF is found, additional investigation and analysis is required to establish item bias.

There can be situations in which differential functioning of a test item operates to distinguish high proficiency students from low proficiency students when such distinction is a construct of the test itself. An accounting exam, for example, may have items involving complex comparison of ledger entries. Test participants with lower skills in solving the test problem will perform more poorly than test participants with more experience and skill. There would be differential results, but the difference is precisely the characteristic that the test was intended to identify. In such situations, DIF would not represent inappropriate bias or a defect in the test.
In contrast, a test item may advantage students of one culture over those of another culture and give the impression that students of a particular culture are more or less proficient than another. A test for English Language Learner (ELL) proficiency in which a test item results in higher performance by Asian students than European students would represent DIF. If the difference can be explained by a test language reference that advantages students of Asian culture, rather than basic language proficiency, the item would be considered biased. When the different performance is a result of cultural context or cues in the test item that advantage a particular native language speaker, the resulting performance difference is problematic. The challenge is in determining the basis for the differential performance when cultural factors can be confounded with proficiency.

A reason for choosing DIF analysis as a framework for this content analysis is that the question respecting application of the assessment tool focuses upon assessments to compare proficiency of students whose native language and culture is different from that of the students for whom the assessment was initially designed. To address this question, the first analytic step would be to demonstrate whether there is DIF with respect to at least some of the test items. Absent DIF for one or more test items, most objections to the cross-cultural and cross-national use of the assessment tool would be defeated.

The body of research regarding the development and refinement of DIF in the area of language testing is described in an article by Ferne and Rupp (2007). Fifteen years of research was reviewed and summarized, and factors that include identification, reporting and drawing of inferences from DIF are discussed. A study of Canadian students that included ELL and native English speakers (L1) used DIF to examine results and test validity. That study found vocabulary in test items advantaged L1 students, while grammatical questions advantaged ELL students in test results (Kim & Jang, 2009). This leads to an additional consideration and potential obstacle to interpretation of DIF analysis, the multidimensional nature of test items impacting determination of existence and basis for differential results. When multiple factors contribute to test performance, control for the relative influence that each factor may have becomes important. In addition, each factor may influence performance of different test populations in different ways.

Snow and Oshima (2009) conducted a study using DIF and attempting to analyze testing that was multidimensional. Their study encountered difficulty separating and identifying the different dimensions scored in the test, as well as measuring the confounding influence of multiple dimensions. The assessment of language and reading comprehension is inherently multidimensional and perhaps more difficult to measure than other content areas or skills. To process meaning from text requires a number of interactive skills. Vocabulary, grammar, background knowledge, and awareness of literary devices all contribute to the derivation of meaning from a reading passage in order to respond to test questions regarding the content and inferences.

Cross-Cultural Assessment Factors

Growing awareness of implications of multicultural education has led to study of cultural bias in assessment tools. Teaching diverse groups of students with respect for their cultural differences
instead of assuming that all must comprehend meanings based upon the dominant culture is viewed as more central to fair and accurate assessment. A study of cross-cultural assessment issues discusses the opinions of professional multicultural educators, and identifies major factors that impact the fairness of test measurement and assessment in cross-cultural settings (Marbley, Bonner, & Berg, 2008). In a slightly different formulation, another study addresses whether a technically valid measurement also is a fair test when the results are used to make inferences across diverse groups (Xi, 2010). Conversely, it is possible for a test that is not deemed biased as a whole (statistically significant bias) may, nevertheless, have items biased against a group of test takers (Takala & Kaftandjieva, 2000). Kunnan (2010) addresses the specific application of this fairness argument to language assessment.

The increasing reliance and intensity of focus upon testing for purposes of allocating educational resources and opportunities applies significant pressures and potential for corresponding harm upon people of color and people whose language and culture differ from the dominant culture upon which the assessment tool is constructed. Assessment tools developed with philosophical underpinnings of assimilation, perhaps unstated, presume that the measure of performance or “proficiency” equals compliance with dominant culture values, norms, and standards, rather than underlying skill.

A test item, for example, may involve describing the proper way to cross a street. One student describes going to the designated crosswalk, waiting for the light, and then crossing to the other side of the street. Another student may describe approaching the curb, looking to see that no traffic is coming from either direction, and then crossing to the other side. Both students have described the target activity functionally, but the first student might be deemed to have given the correct answer based upon implicit cultural norms embedded in the test question or required response. Accordingly, assessment tools must be carefully scrutinized from a multicultural perspective to identify culturally relevant factors that may suggest different meanings and may yield inaccurate and discriminatory assessment results.

A cultural assessment approach must also include the appropriate and responsible use of the dissemination of research and measurement instruments, including their selection, administration, scoring, interpretation, and communication, as well as recognition of the preponderance of bias existing within the practice of the field of educational measurement and other disciplines within the academy (Marbley et al., 2008, p. 12).

In an extensive review of cross-cultural assessment literature, Marbley, Bonner, and Berg (2008), gleaned central information by positing a framework based upon three questions, two of which are more relevant here: (1) What factors in assessment and measurement are influenced or impacted by multiculturalism? and (2) What are professional educator perceptions of the experiences of people of color regarding assessments deemed unsuccessful and those deemed successful? The survey included 14 participants from various disciplines and ethnic backgrounds, all of whom had experience with bias or biased instruments in the field of education.
The survey results suggested four themes to be considered in ethically incorporating and advancing multiculturalism in educational assessment. The observations and conclusions of the study participants were compared with available research regarding multicultural issues in education and counseling and found consistent. **Language bias** was identified as an important theme. The characteristics defining this bias for surveyed educational professionals included not only bilingual and non-native language considerations, but also the “cultural, intonation, semantic, or dialectal” and nuanced meaning differences of native speaking students of color. **Professional bias** was another important theme, the awareness and skill of the educational professional in dealing with students from diverse backgrounds. Failure to perceive differences based upon culture, or using culture against the student can be damaging not only to the validity of assessment results but also on the self-concept and esteem of the student. **Practice bias** is related to professional bias, except that it refers to the institutional norms and practices embodying and reinforcing bias through the implementation of policies and programs. In the present context, practice bias would manifest in the institutional decision to use an assessment tool that is inherently biased or to misapply a tool that is not inherently biased in ways that result in bias through interpretation and inferences. **Instrument bias** refers to the tools used to measure characteristics of competency or performance that disadvantage certain groups of students being assessed because of cultural bias. Although this aspect of assessment can be highly technical in the analysis of the degree of bias, the literature from multiculturalism in counseling reminds us that the very process of testing is interactive. It does not assess the characteristics of inanimate objects, like measuring the weight of a rock. When a student takes a test and fails to comprehend a test question that is obscure because of cultural bias, that failure has an internalized impact on the student and his or her self-concept of academic competency (See, Sackett, Borneman, & Connelly, 2009b)

**Validity and Fairness**

Generally speaking, the debate among researchers concerning validity and fairness of using test results derived from a diverse test group centers upon the distinction between measurement bias and predictive bias. While these two factors can influence each other, they are distinct. Both depend, however, on the construct for which the assessment is used as an evaluation tool. Considering the importance of high stakes testing, Sackett and Borneman examined issues relating to the validity, bias, and fairness of test instruments. Responding to Wichert and Millsap’s (2009) critiques of their analysis, they elaborated upon their prior reference to stereotype threat, a kind of measurement bias in which minority test takers underperform in relation to a perception that they will be lower performers because of their race (Sackett, Borneman, & Connelly, 2009). Their conclusion was that, if stereotype threat resulted in observed underperformance on the test by minorities, then the assessment would under predict the competency of that group with respect to the construct for which the assessment tool was being used.

An additional group of researchers have examined the inferences or use that can be drawn from test results obtained from a diverse population. Examining this question in the context of language testing, Kane (2010) explores the differences and relationship between validity and fairness in use of test results. He notes that the relationship between validity and fairness
depends upon the definition of each term. He uses a broad definition of both validity and fairness, and treats them as separate constructs for purposes of a critique of Xi’s (2010) model for fairness analysis. Xi proposes a model in which different types of inferences or decisions based upon test scores are compared across various groups in the test population. Variability and strength of inferences, or outcomes and consequences, may suggest specific issues regarding the fairness of the assessment tool.

**Cross-National Assessment Implications**

Although there is increasing discussion of globalization and the cross-national comparison of student proficiency and achievement for purposes of education policy, the amount of research and analysis of the effectiveness of assessment tools to make valid comparisons has not received commensurate attention. One critical analysis article questions the implications of standardized testing, as a technological tool, for interpreting the status of public education and making cross-national comparisons. A technological tool developed in, and for, a particular culture may be uncritically applied across cultures with assumptions that it is “objective,” when the assessment tool may have embedded notions of political, ideological, and socio-cultural structure or discourse (Viruru, 2006). The value of inferences drawn from such use is questionable.

Standardized testing is a form of technology. Like any tool, therefore, its utility depends upon the skill and purposes of the agent that wields it. One of the risks in the use of standardized testing is seduction by the idea that the scientific and empirical nature of the tools equates with the objectivity of the tool itself. Viruru (2006) points to a growing body of research suggesting that such testing can be an instrument for imposing political and corporate ideology upon children. To the extent that standardized tests with embedded cultural and ideological bias are exported for use in other countries, such dissemination can be seen as a form of capitalist colonialism or hegemony. The bias of concern is not the specific geopolitical agenda of a particular country, but rather a trans-national corporate agenda (Miyoshi, 1993). Standardized tests developed by major corporations and “testing companies” increase in use, and major reports of the level of educational system proficiency are authored by international business organizations instead of education professionals.

The result of this trend is that governments, including the U.S. government, are now shaping education policy based upon assessments and metrics interpreted in the context of corporate and capitalist paradigms. Educational systems must be more “efficient” and must increase “production” of students with measurable skills to support national goals for “global economic competition.” To what extent, therefore, do these standardized tests measure a particular style of thinking and particular applications of skills? In this perspective, it is perhaps easier to see how tests developed in the United States and exported to other countries to assess the proficiency of children in those countries could reflect a colonial bias. How well students of other countries do on the examinations depends upon how well they emulate and adopt cultural and ideological norms embedded in the standardized test. How the students of another country rank comparatively for purposes of supporting global economic competition requires that one first accept that a central purpose of education is to support embedded global economic competition.
values. Moreover, it assumes that national strategy for development of human potential in the countries to which the testing is exported adopt the model and rules established by the trans-national and multinational corporate organizations. Use of such testing without consideration of the cross-cultural implications may lead to interpretation of results and adjustments to curriculum that are not in the best interests of the nation or the children being educated.

Content Analysis

Methodology

The Fountas & Pinnell Benchmark Assessment is structured in a hierarchy of reading assessment exercises starting with level A, as the least difficult, through level Z, the most challenging. For each level, a prescribed reading “text” is provided which is delivered either aloud to the student, read by the student, or a combination of the two. Following delivery of the textual passage, the student responds to a set of questions regarding content and meaning of the text passage. Generally, half the questions seek a response based upon the student’s knowledge and comprehension of the literal content, recognition of specific facts or events described in the text. The other half of the questions seek responses based upon inferences drawn from the facts, events, or interactions, and responses of characters described in the text materials. If a student answers all questions correctly, the student may proceed to the next higher assessment level. If a student answers one question incorrectly, the level is considered appropriate for the student. If the student gets two or more questions wrong, the level is considered too difficult and the student is retested at a lower level.

For purposes of the content analysis, evaluation of both the text passages and the questions was relevant because the performance on the assessment, and subsequent reading comprehension level designation, is determined by the ability of the student to both recognize the subject matter of the passage and to derive inferential meaning from what is described. For this purpose, the sample consisted of alternate level exercises, levels B-D-F, etc. through level Z.

The terms in these selected level items, other than conjunctions and prepositions, were analyzed for classification into three different groups for purposes of their potential differential function. More comprehensive analysis could include all test items and include prepositions which may have culturally bound meaning in some languages. The first category was the neutral group, neither bound to nor differentiated by any cultural reference or norm. Typical usage of terms like “house” and “walk”, for example, would not be subject to clearly different meanings regardless of the background or culture of the student.

The second category included terms subject to multiple or alternative interpretations that could be culture bound, but were accompanied by text that defined or clarified the reference, contextualizing it. An example of this category would be “tack” used as a verb. Without additional language to clarify the intended meaning, the reader would not know whether the word indicated attaching something, or changing direction when sailing a ship. In addition, students from lower socio-economic backgrounds or cultures not typically associated with water transportation would
be unlikely to know about the latter meaning. If surrounding language in the passage describes sailing and the direction of the boat, this term would be placed in the second category for purposes of analysis.

The third category included terms subject to alternative interpretations that could be culture bound, but which were non-contextual, that is, not supported by language that clearly defined or clarified the intended meaning. In addition, terms that have specific cultural meaning for certain cultures that would be different from the dominant cultural meaning would fall in the third category. An example of this last type of term might be “uptight” that means a state of being nervous and insecure in dominant culture, but in African American culture means being comfortable and pleased with the state of things. So, if a sentence said only that John was feeling “uptight,” an African American student may give an incorrect response when asked to describe how John was feeling.

These 13 exercises were then coded to reflect two dimensions that might indicate a difference in function [DIF] of the item. First, the terms were coded into categories, as described above. Next, the determination was made whether the questions to which the student responded required the student to clearly understand the term that had been identified. Each question in the assessment has a target or “appropriate” response that is provided by Fountas and Pinnell. In some instances, the question did not require the student to know the meaning of the term in order to respond. If the text, for example, stated that “When he woke in the morning, John was feeling uptight,” and the question asked when was John feeling differently, the understanding of “uptight” would not be required to answer the question in the manner the assessment targeted. In contrast, if the response to the question was dependent upon a clear understanding of the culture bound term, that condition was coded. The resultant analysis could then show the number of culture bound terms that were or were not clarified, and the relevance of those terms to the required response in the assessment.

Results

In the text passages for the level exercises reviewed for purposes of the content analysis, there were a total of 499 sentences. Each sentence was analyzed and coded as described above. Sentences with only terms in the first category were not referenced in the analysis because no term in the sentence was deemed to potentially cause different function in the student performance on the assessment. The total number of sentences out of 499 containing terms that were in the second category was 11 (see Figure 1). The sentences out of 499 that contained terms coded in the third category numbered 26 (see Figure 2). Of the questions in the assessment dependent on understanding of the coded term, only two of the literal questions were associated. Of the questions in the assessment dependent on understanding of the coded term, four of the inferential questions were associated (see Figure 3).
Each instance of a problematic term in a sentence creates potential for misunderstanding by the student of the content of the reading passage or the questions about the passage. The basis for the misunderstanding and potential “incorrect” response would be a factor that should be irrelevant to the assessment, i.e. the cultural background of the student. Where the cultural or language difference impacts the student response, there is potential differential functioning of the test item. Since getting one question wrong or two questions wrong, may make the difference in the level designation of the student, the sensitivity threshold is fairly low for this particular assessment.

It may be that the relative impact of a question involving a second category problem is lower because the information is available within that text for the student to resolve confusion. In questions that involve problems in the third category, there is no similar contextual support to clarify the meaning and help the student find the target response. Moreover, the sentences with problems in the third category were associated with inferential questions that are more dependent upon the processing of meaning than upon simple vocabulary identification.
Conclusion

From this preliminary content analysis, a few tentative conclusions seem appropriate. A complete analysis of the Benchmark Assessment would be helpful to evaluate the assessment tool, since there are multiple sets of reading passages and questions. The variability of the amount of reading text per item also merits examination. Supporting information, particularly with regard to the target or “appropriate” responses should be examined, and teachers using the system should be advised of a potential range of responses that are more culturally sensitive and may be acceptable in designating reading comprehension levels. Where assessment passages and questions contain terms that are clearly culture bound and unexplained, such questions should be removed and substituted.

In general, the Fountas and Pinnell system could be a useful tool for assessing basic reading and literacy proficiency for students from the dominant culture in the United States. Care should be taken when applying the system to non-dominant culture groups of students in culturally diverse populations within the United States. Use of the system without careful attention to these culture bound issues for students in countries other than the United States should be approached with substantial caution. Marketing of the system for international use without such caveat could be substantially misleading.

References


