Paper Title Tests of Alignment Among Assessment, Standards, and Instruction Using Generalized Linear Model Regression

Author(s) Gavin W. Fulmer, National Institute of Education - Nanyang Technological University

Session Title Using Assessment Data to Improve Curriculum and Instruction

Session Type Roundtable Presentation

Presentation Date 5/1/2013

Presentation Location San Francisco, California

Descriptors Assessment, Data Analysis, Research Methodology

Methodology Quantitative

Unit Division H - Research, Evaluation and Assessment in Schools

Tests of Alignment among Assessment, Standards, and Instruction Using Generalized Linear Model Regression

Gavin W. Fulmer

National Institute of Education (Singapore)

1 Nanyang Walk, Singapore 637616

Paper presented at the 2013 annual meeting of the American Educational Research Association, San Francisco, CA, USA.

Abstract

An essential aspect of the current climate of accountability is to ensure that assessments are well-aligned with the standards or curriculum they are intended to measure. However, there is relatively little prior study on how alignment or related discrepancies among source documents are to be interpreted. The present paper explores analyses of alignment and marginal discrepancies as a special case of the general linear model (GLM). A general approach for such analyses is suggested, and an example is given using both traditional alignment and GLM regression analyses. Results show that the use of GLM allows more rigorous interpretation of discrepancies between source documents than alignment analysis alone, including determination of whether discrepancies are significant or not.

Keywords: alignment; standards; standardized testing; general linear models

Tests of Alignment among Assessment, Standards, and Instruction Using Generalized Linear Model Regression

The momentum for systems of educational accountability continues at the state, national, and international levels (e.g., Elstad, Turmo, & Guttersrud, 2011; Jaafar, 2011; Mattei, 2012; Müller & Hernández, 2010; Ng, 2010). To achieve the purported goal of accountability, the assessments that are to be implemented must be valid and representative to allow policymakers, educators, researchers, and the public to understand the extent to which students and schools are meeting expectations (Beck, 2007; D'Agostino, Welsh, & Corson, 2007; Rothman, Slattery, Vranek, & Resnick, 2002). An essential aspect of this need is to ensure that assessments are well-aligned with the standards or curriculum they are intended to measure (Bhola, Impara, & Buckendahl, 2003; Polikoff, Porter, & Smithson, 2011; Porter, 2002).

However, while alignment is an important requirement for the development and interpretation of standardized tests, there is relatively little prior study on how alignment indices or the related discrepancies among source documents are to be interpreted. That is, while an analysis may result in some estimate of test-standard alignment, little prior research has explored how to test whether observed alignments or discrepancies are statistically significant. In previous work, the author and colleagues addressed statistical tests for the alignment index using simulation analyses (Fulmer, 2011; Polikoff & Fulmer, in press).

While these previous papers have demonstrated numerical methods for simulating alignment indices to estimate the respective critical values, such methods also treat the alignment index as if it were a random, continuous variable. However, because the alignment index is based on coding documents into fixed categories, the underlying data from which alignment indices are calculated is categorical. To address this discrepancy, the present paper explores analyses of marginal discrepancies in alignment studies as a special case of the general linear model (GLM). The purpose of the study is to suggest a more general approach for estimating whether there are significant differences in alignment among tests, standards, and instruction.

Literature Review

This study builds on previous work on alignment among assessments, instruction, and standards or curriculum. Approaches to calculating and interpreting alignment are varied, such as the *Depth of Knowledge* framework proposed by Webb (2007), or the method described by Porter (2002) and used for the Surveys of Enacted Curriculum (SEC; Council of Chief State School Officers, 2004). Porter's (2002) alignment index is the focus of the current paper. Porter's alignment index is easily calculable and is widely known and used for policy-related analyses such as the SEC (Council of Chief State School Officers, 2004; Polikoff, et al., 2011; Porter, Smithson, Blank, & Zeidner, 2007). Furthermore, Porter's alignment index has been and can be applied to any combination of assessments, instruction, and curriculum (Liang & Yuan, 2008; Liu & Fulmer, 2008; Porter, Polikoff, Zeidner, & Smithson, 2008). For the sake of simplicity, the remainder of this paper will use the term *curriculum*—except in describing studies that focus explicitly on standards—while recognizing that studies of alignment may have different foci if applied to standards documents rather than curriculum or to enacted rather than mandated curriculum.

Prior research has demonstrated that the degree of alignment among tests, standards, and instruction can vary considerably (e.g., Liu & Fulmer, 2008; Rothman, 2003) although in unexpected ways. For example, Porter (2002) found that there was approximately equivalent alignment of between each state's tests and standards as across states. Rothman and colleagues

(Rothman, et al., 2002) found that, while individual items or groups of items may align well to a set of standards, a test overall may overemphasize or underemphasize particular content areas or skills. Similarly, Polikoff and colleagues (Polikoff, et al., 2011) argued that test coverage was only marginally acceptable if state tests are to be used for high-stakes decisions such as student advancement or educator evaluation (particularly under the value-added modeling approach, e.g., Amrein-Beardsley, 2008).

Additionally, Porter's alignment has been applied to particular subfields of education, such as science education. Liu and Fulmer (2008) calculated the alignment between New York State Regents physics and chemistry exams and the respective standards, showing that there is noticeable differences in alignment indices over time for the same testing program and subject matter. In another area of work, Liang and Yuan (2008) and Liu and colleagues (2009) examine alignment among standards and exams in China, the US, and Singapore, and find important discrepancies in the level of cognitive complexity that the exams measure compared to the respective curriculum and standards. In their findings, Chinese and Singaporean curriculum materials require lower-level cognitive skills than their standardized tests, whereas this discrepancy is much smaller or non-existent for the US standardized tests.

From a methodological perspective, prior work has examined the alignment concept as a psychometric quality of a test (e.g., Beck, 2007; Martineau, Paek, Keene, & Hirsch, 2007), or as a teacher-level variable (Porter, et al., 2007). However, only relatively recently has there been work on the extent to which an observed alignment can be considered “high” or “low,” based on aspects of the coding process and coding assumptions (Fulmer, 2011). This has been corroborated with later study that extended that simulation method to coding conditions typical for the SEC, such as the complexity of the coding scheme or the number of raters involved

(Polikoff & Fulmer, in press). However, these prior articles still draw upon a simulation algorithm. That is, the methods described can only provide an estimate of the significance of an alignment index based on the range of values that could occur by chance, given the coding conditions. Furthermore, these approaches involve the assumption that the alignment index can be treated as a continuous random variable, whereas it is calculated from categorical data based on raters' analyses of documents (whether standards, curriculum, or test items) or on teachers' responses to Likert-type survey questions. Thus, prior work has not considered the categorical nature of the coding scheme involved. To address these issues, the present paper presents a basic overview and demonstrates the use of a generalized linear model for categorical data that can be used to analyze alignment among tests and curriculum.

Methods

This study presents a basic summary of the method for use of the generalized linear model (GLM), and then demonstrates the findings of that GLM method with two sample datasets. The sections below present a description of the Porter alignment index and compare that with the GLM approach. The following sections describe the context for the sample data used in the study and the analyses undertaken here.

Calculation of the Porter Alignment Index

Under the Porter alignment index approach, any pair of documents—a test and the associated standards, for example—are compared by first coding each document according to two categorical variables. The categorical variables could be any variable of theoretical or practical importance. Prior research has examined test items and standards statements by content area (e.g., scientific topics, English language content) and by cognitive demands (e.g., recollection or comprehension under Bloom’s taxonomy). This process results in two tables, one for each document, and an alignment index is calculated (e.g., Fulmer, 2011).

Sample Data

As a demonstration of the approach, data for the paper are drawn from an analysis of New York state’s Regents Exams by Liu and Fulmer (2008). The data set has two variables corresponding to the coding dimensions, a variable for the document source, and a variable containing the frequencies. Two *documents* were coded: a state physics test (document 1) and the respective physics standards (document 2). Both documents were coded on two dimensions: *content*, the physics content of the test items and curriculum statements; and *cognitive demand*, the cognitive activity indicated for the test items or curriculum statements according to Bloom’s taxonomy. For each level of the content and cognitive demand, there was a frequency of points associated with the respective document. Thus, the example study consisted of four variables: document, content, cognitive demand, and frequency. The coding results can be presented as a three-way contingency table (Table 1). The four variables identified are similar to other studies based on Porter’s alignment approach (e.g., Liang & Yuan, 2008; Polikoff, et al., 2011). These data are analyzed via the standard alignment approach, as well as using a GLM.

Data Analyses

The alignment index and data on marginal discrepancies between the test and the standards are calculated following the approach of Porter (2002), and the alignment index is tested for statistical significance as described by Fulmer (2011). As an alternative to the alignment index approach, the frequencies produced in the tables can be analyzed using a generalized linear model (GLM). GLMs extend regular, multiple regression models by allowing analyses of data that do not follow a strict normal distribution (see Nelder & Wedderburn, 1972, for the original formulation of GLMs). This does not require the calculation of an alignment index. Rather, one can analyze whether there is a statistically significant difference in the probability of the observed ratings. Furthermore, it is flexible to nonparametric models, such as analyses of the contingency tables produced in alignment studies.

When estimating a GLM for alignment purposes, the process begins by creating a data set of the coding frequency for each of the documents. After forming the data set, one tests a series of general linear models to identify whether there is statistically significant dependence among the observed frequencies for the Source (document). Because the data are observed frequencies from raters, and the mean frequencies in the cells tend to be relatively small (particularly for cases such as the SEC tables), the most appropriate distribution is the Poisson distribution (Nikoloulopoulos & Karlis, 2008) rather than the logistic distribution (cf. Hosmer & Lemeshow, 2000). The GLM procedure is also flexible to handling data where individual cells have value of zero.

A null model for the analyses can be estimated by fitting a model with only an intercept term, but the fully independent model is recommended (Faraway, 2006). Using Faraway's (2006) notation, let p_{ijk} be the probability that an observation falls into the cell of stratum i (e.g., Source) at row j (e.g., Content) and column k (e.g., Cognitive Level). Furthermore, let p_j be the marginal probability that the observation is in row j (for any stratum and any column), and likewise for p_i and p_k. When the probabilities are independent (but not mutually exclusive), the probability of an observation in any cell is then $p_{ijk} = p_i \cdot p_j \cdot p_k$. Thus, the expected value for Y_{ijk} is $n \cdot p_{ijk}$, and $\log EY_{ijk} = \log n + \log p_i + \log p_j + \log p_k$. This is model of mutual independence can be parameterized in the following way:

$$(\text{Freq})_{ijk} = \beta_0 + \beta_1(\text{Source}) + \beta_2(\text{CogLevel}) + \beta_3(\text{Content})$$

This model can be compared with models of joint dependence, such as:

$$(\mathrm{Freq})_{ijk} = \beta_0 + \beta_1(\mathrm{Source}) + \beta_2(\mathrm{CogLevel}) + \beta_3(\mathrm{Content}) + \beta_4(\mathrm{Source}{\times}\mathrm{Content}) + \beta_5(\mathrm{Source}{\times}\mathrm{CogLevel})$$

It is also typical to estimate a comparison model with the three-way interaction term. However, this model is saturated and it must be interpreted with caution.

GLM regression models are compared using *deviance*, similar to residual variance in analysis of variance (ANOVA). Furthermore, as with other statistical modeling techniques, two or more models can be compared by examining their relative fit to the data, with adjustment for the number of predictors included, such as AIC (i.e., Akaike's Information Criterion). The AIC are compared by estimating each of the models, and identifying the model with the lowest AIC.

Results

Results from the GLM analyses for are shown in Table 2. The model-fitting results will be discussed first, then the results of particular terms. Model 1 was the fully saturated model, appropriate for comparisons, but cannot be analyzed further. It has 0 residual deviance, 0 residual df, and the highest AIC (215.37). Model 2 removes the three-way interaction but retains all two-way interactions to test joint dependence among pairs of the three variables. This paper focuses on possible effects of Source document, so Model 3 removes the joint dependence term for Cognitive Level and Content. Finally, Model 4 is the fully independent model, without joint dependence terms for Source—with Cognitive Level and with Content.

Model 3 has the lowest AIC of the estimated models (154.61). Furthermore, likelihood-ratio tests for the models show that the increase in residual deviance was significant for models 3 and 4 (χ^2=19.35, df = 9, p<.05). Thus, model 3 is preferred as having superior model-data fit.

This shows that the superior model has joint dependence of Source with Content and with Cognitive Level.

Now consider the model terms. There are significant main effects for both Cognitive Level and Content, but non-significant effect for Source document. This indicates that there are differences in the distribution of frequencies according to each of Cognitive Level and Content, when analyzed alone, but no difference between Sources. This is not unexpected. There is also statistically significant interaction between Source and Cognitive Level. This indicates a significant difference in the distribution of marginal frequencies for Cognitive Level between the two Sources.

These findings can then be compared with a similar analysis of the alignment index between the test and standards. Calculation yields an alignment index of 0.80. Using the method proposed in Fulmer's (2011) simulation study, this index is statistically significantly different from what alignment index could occur by chance (0.689), equivalent to a z-score of 2.56 ($p<.05$). In addition, one can examine discrepancies between the proportions of test items and standards statements by the content (Figure 2) or cognitive level (Figure 3).

In a typical alignment analysis, one could only discuss the alignment and the discrepancies descriptively, but drawing upon the GLM results it is possible to argue that there is not a significant difference between the source documents by content area, but that there is a significant difference between the source documents by cognitive level (Figure 2).

Discussion and Implications

While the continued emphasis on school accountability based on standardized tests has both champions and detractors (cf. Wiliam, 2010), it is undeniable that test-based accountability will continue to be influential for policymakers, researchers, school personnel, and others.

Alignment among tests, standards, and instruction is a significant requirement for valid interpretation of standardized test results. As efforts continue to increase the level of alignment among tests, instruction, and standards, it is also necessary to develop further the field’s ability to understand and interpret alignment correctly. However, previous analyses of alignment have been unable to examine the statistical significance of observed alignment indices or of discrepancies between observed source documents.

The application of GLM to detect if there are differences between tables of coded documents—such as tests and standards—allows examination of differences between the documents that goes beyond that which is available via Porter’s (Porter, 2002) alignment index approach. The present paper describes an approach using the general linear model (GLM) to estimate whether any pair of documents—such as a test and the associated standards—have statistically significant differences in their ratings. This approach does not replace prior work on estimates of alignment (Fulmer, 2011; Porter, 2002); rather, it provides an additional method to test whether observed differences in ratings are statistically significant. This has potential to increase the field’s use of indices and other approaches to examine extent of alignment and sources of discrepancy and misalignment. Additionally, future research can extend the present study to include analysis of multiple rater effects or other variations that reflect differences in practice among coding schemes (Council of Chief State School Officers, 2004).

References

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher, 37*(2), 65-75.

Beck, M. D. (2007). Review and other views: "Alignment" as a psychometric issue. *Applied Measurement in Education, 20*(1), 127-135. doi: 10.1207/s15324818ame2001_7

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. . *Educational Measurement: Issues and Practice*(Fall), 21-29.

Council of Chief State School Officers. (2004). Surveys of enacted curriculum. Madison, WI: Wisconsin Center for Education Research.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1-22. doi: 10.1207/s15326977ea1201_1

Elstad, E., Turmo, A., & Guttersrud, Ø. (2011). Problems induced by amalgamation of pedagogical progressivism and educational accountability: Oral exams with prior preparation time in Norwegian secondary schools. *Problems of Education in the 21st Century, 30,* 22-34.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects, and Nonparametric Regression Models*. New York: Chapman & Hall.

Fulmer, G. W. (2011). Estimating critical values for strength of alignment among curriculum, assessments, and instruction. *Journal of Educational and Behavioral Statistics, 36*(3), 381-402.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.

Jaafar, S. B. (2011). Performance-based accountability in Qatar: a state in progress. *Compare: A Journal of Comparative & International Education*, *41*(5), 597-614. doi: 10.1080/03057925.2011.555139

Liang, L. L., & Yuan, H. (2008). Examining the alignment of Chinese national physics curriculum guidelines and 12th-grade exit examinations: A case study. *International Journal of Science Education, 30*(13), 1823-1835. doi: 10.1080/09500690701689766

Liu, X., & Fulmer, G. W. (2008). Alignment between the science curriculum and assessment in selected NY State Regents Exams. *Journal of Science Education & Technology, 17*(4), 373-383. doi: 10.1007/s10956-008-9107-5

Liu, X., Zhang, B., Liang, L. L., Fulmer, G. W., Kim, B., & Yuan, H. (2009). Alignment between the physics content standard and the standardized test: A comparison among the United States-New York State, Singapore, and China-Jiangsu. *Science Education, 93*(5), 777-797.

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, Comprehensive Alignment as a Foundation for Measuring Student Progress. *Educational Measurement: Issues & Practice, 26*(1), 28-35. doi: 10.1111/j.1745-3992.2007.00086.x

Mattei, P. (2012). Market accountability in schools: policy reforms in England, Germany, France and Italy. *Oxford Review of Education, 38*(3), 247-266. doi: 10.1080/03054985.2012.689694

Müller, J., & Hernández, F. (2010). On the geography of accountability: Comparative analysis of teachers' experiences across seven European countries. *Journal of Educational Change, 11*(4), 307-322. doi: 10.1007/s10833-009-9126-x

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, 135*(3), 370-384.

Ng, P. T. (2010). The evolution and nature of school accountability in the Singapore education system. *Educational Assessment, Evaluation & Accountability, 22*(4), 275-292. doi: 10.1007/s11092-010-9105-z

Nikoloulopoulos, A. K., & Karlis, D. (2008). On modeling count data: a comparison of some well-known discrete distributions. *Journal of Statistical Computation & Simulation, 78*(3), 437-457. doi: 10.1080/10629360601010760

Polikoff, M. S., & Fulmer, G. W. (in press). Refining methods for estimating critical values for an alignment index. *Journal of Research on Educational Effectiveness*.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*(4), 965-995. doi: 10.3102/0002831211410684

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and content standards. *Educational Measurement: Issues & Practice, 27*(4), 2-14. doi: 10.1111/j.1745-3992.2008.00134.x

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a Teacher Variable. *Applied Measurement in Education, 20*(1), 27-51. doi: 10.1207/s15324818ame2001_3

Rothman, R. (2003). Imperfect matches: The alignment of standards and tests. Washington, DC: National Research Council.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). Benchmarking and alignment of standards and testing. Los Angeles, CA: Center for the Study of Evaluation.

Webb, N. L. (2007). Issues Related to Judging the Alignment of Curriculum Standards and Assessments. *Applied Measurement in Education*, *20*(1), 7-25. doi: 10.1207/s15324818ame2001_2

Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*, *45*(2), 107-122. doi: 10.1080/00461521003703060

Tables

Table 1

Three-way contingency table for frequencies by Content Topic and Cognitive Level for two Source Documents

	Content Topics				
Cognitive Level	Electricity	Energy	Motion & Forces	Properties of Matter	Waves
Source Document: Curriculum					
Remember	0	0	0	0	0
Understand	7	8	13	9	11
Apply	7	7	18	3	9
Analyze	2	2	1	0	2
Evaluate	0	0	0	0	0
Create	1	0	0	0	0
Source Document: Test					
Remember	0	0	1	1	1
Understand	7	3	10	2	10
Apply	6	11	19	3	11
Analyze	0	0	0	0	0
Evaluate	0	0	0	0	0
Create	0	0	0	0	0

Note. Data for the frequencies in the table are drawn from Liu and Fulmer’s (2008) analysis of New York State Regents physics exams and curriculum.

Table 2

GLM Analysis Results for Four Nested Models

		Model			
Source	df	1	2	3	4
CogLevel	5	320.52 ***	320.52 ***	320.52 ***	320.52 ***
Content	4	29.77 ***	29.77 ***	29.77 ***	29.77 ***
Source	1	1.22	1.22	1.22	1.22
Source×CogLevel	5	17.64 **	17.64 **	17.64 **	
Source×Content	4	2.12	2.12	1.71	
CogLevel×Content	20	15.73	15.73		
CogLevel×Content×Source	20	3.09			
Model Residual Deviance		0	3.09	19.23	38.58
Model Residual df		0	20	40	49
AIC		215.37	178.47	154.61	155.95

Figures

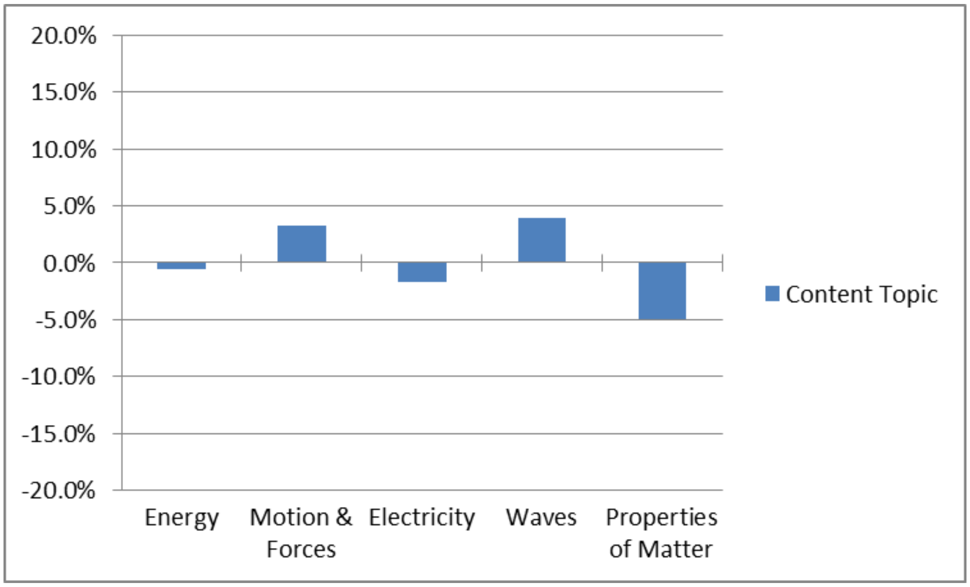

Figure 1. Chart showing marginal discrepancies for content areas between the test and the standards. Positive values indicate the test shows greater emphasis than the standards on the respective content area.

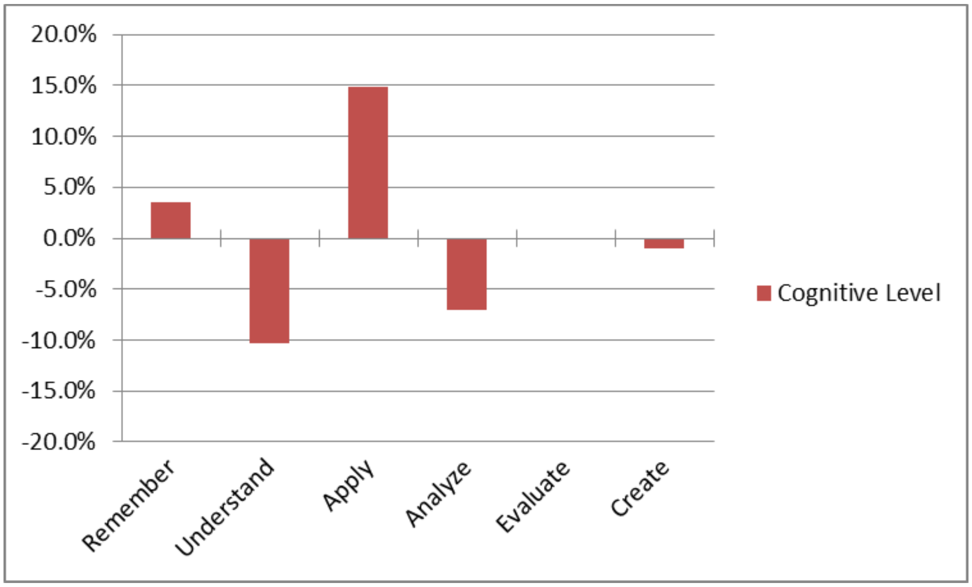

Figure 2. Chart showing marginal discrepancies for cognitive demands between the test and the standards. Positive values indicate the test shows greater emphasis than the standards on the respective cognitive level.