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

[in-press in *Educational Assessment, Evaluation, and Accountability*]

Gavin W. Fulmer

National Institute of Education (Singapore)

1 Nanyang Walk

Singapore 637616

Morgan S. Polikoff

University of Southern California

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

Abstract

An essential component in school accountability efforts is for assessments to be well-aligned with the standards or curriculum they are intended to measure. However, relatively little prior research has explored methods to determine statistical significance of alignment or misalignment. This study explores analyses of alignment as a special case of the generalized linear model (GLM). A general approach for such analyses is suggested, and examples are given of analyses with traditional alignment and GLM regression using data from two previously published studies. Results from the GLM are compared with ordinary least-squares (OLS) regression. Findings show that the GLM method allows more informative analysis of differences between source documents than alignment indices alone, including determination of whether marginal discrepancies are statistically significant or not.

Keywords: alignment; curriculum; educational assessment; standards; standardized testing; generalized linear models

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

The momentum towards accountability through standardized tests 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 foundation of this accountability effort is that the 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 (cf. Fulmer, 2011). That is, while an analysis may result in some estimate of test-standard alignment, little prior research has explored how to determine whether observed alignments or discrepancies are statistically significant. While recent simulation studies have demonstrated numerical methods for simulating alignment indices to estimate the respective critical values (Fulmer, 2011; Polikoff & Fulmer, 2013), these methods also neglect the underlying discrete nature of alignment data. That is, studying alignment is based on raters' coding of documents into one or more categorical variables and the frequencies of such codes are then used for subsequent analyses. Furthermore,

alignment indices have a fixed range (e.g., 0-1) and do not necessarily follow a normal distribution, so the statistics do not follow the typical assumptions for ordinary regression techniques and parametric hypothesis testing. To address this discrepancy, the present article explores analyses of marginal discrepancies in alignment studies as a special case of the generalized linear model (GLM). The purpose of the study is to demonstrate 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 alignment index 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, as it is easily calculable, 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, curriculum, and instruction (Liang & Yuan, 2008; Liu & Fulmer, 2008; Martone & Sireci, 2009; 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's (2002) study of multiple states found that there was approximately equivalent alignment between each state's tests and the respective standards. 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 subject matter topics or skills. Similarly, Polikoff and colleagues (Polikoff, et al., 2011) argued that tests and standards were not adequately aligned 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 tests and the respective standards, showing that there are 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) examined alignment among standards and tests in China, the U.S., and Singapore, and found important discrepancies in the level of cognitive complexity that the tests measured compared to the respective curriculum and standards. In their findings, Chinese and Singaporean curriculum materials required lower-level cognitive skills than their standardized tests, whereas this discrepancy was much smaller or non-existent for the U.S. 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; Polikoff, 2012a, 2012b). 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 further pursued by a subsequent study that applied the 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, 2013).

While these prior articles are informative, they are still limited by drawing 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 assume that the alignment index can be treated as a continuous random variable, and that the observed and simulated values can be converted to z-scores for identifying critical values. However, each alignment index 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, and has not used methods specifically designed for analyzing categorical data. 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 data sets. 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 variables of theoretical or practical importance. Prior research has examined test items and standards statements by subject matter topic (e.g., scientific topics, English language skills) and by cognitive demands (e.g., recollection, comprehension, etc. according to Bloom's taxonomy). This process results in two tables, one for each document, consisting of the frequency of test points or standards statements in each cell. An alignment index is then calculated based on the absolute discrepancies between the respective cells of each table (see Porter [2002] or Fulmer [2011] for more information and examples of this calculation), using the following formula.

$$P = 1 - \sum \left| A_{jk} - B_{jk} \right| / 2$$

where A_{jk} and Bjk are the proportion of points in the cell at row j, column k of tables A and B, respectively. The index ranges from 0 (no alignment) to 1 (perfect alignment), and is often interpreted as the proportion of content in common between the two sources.

Basic Concept of the GLM

GLMs extend ordinary 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). Under ordinary linear regression, an ordinary least-squares approach is used to estimate parameters that best fit the proposed model to the observed data (e.g., Cohen, Cohen, West, & Aiken, 2002), with the assumption that the data for the dependent variable are continuous and normally-distributed, and that the independent variables are either normally-distributed (in the case of continuous data) or coded to highlight particular effects (either contrast or dummy

coding). By contrast, the GLM approach allows estimation of data where the dependent variables are not continuous or do not follow the typical normal distribution. For instance, GLM regression allows analyses of data that have distributions that follow binomial distributions for data from independent yes/no trials (e.g., flipping coins), or the more general Poisson distribution for frequency data such as raters' codes for alignment studies (Nikoloulopoulos & Karlis, 2008).

Sample Data

Data set 1. The first data set used for demonstration of the approach is drawn from a content analysis of New York State's Regents Exams and associated standards (Liu & Fulmer, 2008). Two *source documents* were coded: a state physics test (document 1) and the respective physics standards (document 2). Both documents were coded on two dimensions: *topic*, the physics subject matter 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 topic and cognitive demand, there was a frequency of points associated with the respective document. Thus, the example study consisted of four variables: document *source*, *topic*, *cognitive level*, 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).

Data set 2. Data for data set 2 come from an alignment study by Polikoff and colleagues (Polikoff, et al., 2011) that applied the SEC alignment coding framework. That data set presents the comparison of each state's standards with the respective standardized tests. As the current paper focuses on demonstrating the use and interpretation of the GLM method—rather than to replicate all extant state-by-state comparisons—only data for the Virginia 6th grade English

Language Arts (ELA) standards and assessments are selected (Virginia Department of Education, 2012), but the approach could be applied to other states and to other tests. As with data set 1, the data are codes for two *source documents*: a test and its associated standards.

The SEC analyses on the Virginia ELA standards use tables with notably greater dimensions than that presented for the NY Regents physics exams. So, there are 63 topic areas coded, and 5 cognitive levels. The SEC approach typically uses multiple raters and allows raters to place test items or content standards into more than one cell of the table. For the sake of simplicity in the demonstration of the method, only the first cell into which an item or statement was assigned is counted, and all of the three raters' codes were counted together in the frequency data (instead of converting to proportions). The coding results can be presented as a three-way contingency table, but for space reasons this cannot be presented here.

Design and Procedure

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). All GLM analyses for this article are conducted in the R statistical analysis environment (Ihaka & Gentleman, 1996). The GLM approach differs from alignment index approaches as it is a complementary approach that does not require calculation of Porter's index. Rather, the analyses test 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, the researcher tests a series of generalized 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. The procedure does not operate well when an entire row or column of data consists of zeroes in both source documents (e.g., when the cognitive level of "Create" has zero test items and zero standards statements associated with it).

The procedure for model comparison used in this article begins with estimation of a fully-saturated model for comparison purposes (Faraway, 2006). The fully-saturated model consists of all main effects and interaction terms. For the present data sets, which consist of independent variables Source, CogLevel, and Topic, the fully saturated model would include all terms including main effect, two-way interaction, and the three-way interaction, as shown in Equation 1. Because all of the possible interactions are included, the fully-saturated model has 0 residual degrees of freedom, and so its term must be interpreted with caution (Ai & Norton, 2003; Faraway, 2006). However, it does provide a useful point of comparison for relative data-model fit of subsequent models in GLM.

$$\begin{aligned}(\text{Freq}) = {} & \beta_0 + \beta_1(\text{Source}) + \beta_2(\text{CogLevel}) + \beta_3(\text{Topic}) + \beta_4(\text{Source} \times \text{Topic}) \\ & + \beta_5(\text{Source} \times \text{CogLevel}) + \beta_6(\text{Topic} \times \text{CogLevel}) \\ & + \beta_7(\text{Source} \times \text{CogLevel} \times \text{Topic})\end{aligned} \qquad \text{(Eqn. 1)}$$

In the next step, the three-way interaction is removed to create a model of joint dependence among the main effects. This estimates the interactive effects among the independent variables in predicting the frequency of points in each cell. For the present sample with three independent variables, the model of joint dependence among all main effects is shown in Equation 2.

$$\begin{aligned}(\text{Freq}) = \beta_0 + \beta_1(\text{Source}) + \beta_2(\text{CogLevel}) + \beta_3(\text{Topic}) + \beta_4(\text{Source} \times \text{Topic}) \\ + \beta_5(\text{Source} \times \text{CogLevel}) + \beta_6(\text{Topic} \times \text{CogLevel})\end{aligned} \quad \text{(Eqn. 2)}$$

For the application to alignment analyses, the focus of study is typically the differences between source documents on the frequency of points related to Topic or CogLevel. So, the model of joint dependence can be reduced further by eliminating the interaction terms that do not contain the Source variable. This model of joint dependence with Source can be parameterized as in Equation 3.

$$\begin{aligned}(\text{Freq}) = \beta_0 + \beta_1(\text{Source}) + \beta_2(\text{CogLevel}) + \beta_3(\text{Topic}) + \beta_4(\text{Source} \times \text{Topic}) \\ + \beta_5(\text{Source} \times \text{CogLevel})\end{aligned} \quad \text{(Eqn. 3)}$$

Lastly, to compare if the joint dependence of source with cognitive level and with topic contribute to the statistical model, one can also examine a model of mutual independence, in which the frequency of points assigned to cell are completely independent of document source. Equation 4 shows this parameterized model.

$$(\text{Freq}) = \beta_0 + \beta_1(\text{Source}) + \beta_2(\text{CogLevel}) + \beta_3(\text{Topic}) \quad \text{(Eqn. 4)}$$

Drawing on this model comparison and testing approach as suggested by Faraway (2006), this article compares multiple nested models consistent with Equations 1 through 4. Figure 1 presents a concise list of the set of generalized linear models that are estimated. Based on the concept of parsimony, each data set is first analyzed using a fully-saturated model followed by

nested models that contain successively fewer terms, ultimately comparing with the mutual independence model.

GLM regression models can be compared in a variety of ways. One such way is through *deviance*, which is similar to residual variance in analysis of variance (ANOVA) but based on maximum likelihood estimates (Cohen, Cohen, West, & Aiken, 2002). Models with lower deviance are considered better fitting to the data. Multiple nested models can be compared using a likelihood ratio test to determine if a change in the model terms results in a significantly better or worse residual deviance. Suppose a model with k terms and deviance D_k is to be compared with a model with $k\text{-}1$ terms and deviance $D_{(k-1)}$. The likelihood ratio test for deviance is performed by calculating $D_{(k-1)}\text{-}D_k$, and comparing this value against a chi-square distribution with 1 degree of freedom (Cohen, Cohen, West, & Aiken, 2002, p. 507). Another way to compare two or more models is by examining their relative fit to the data with adjustment for the number of independent variables included. For the present article, the models are compared on quality of fit using the AIC (i.e., Akaike's Information Criterion). The AIC values are compared by estimating each of the models, and identifying the model with the lowest AIC. This fits with the goals of identifying GLM models that balance data-model fit with parsimony. All models were also compared using BIC (Bayesian Information Criterion) and chi-square, alternatives to AIC; all results were substantively the same regardless of model comparison technique.

While GLM models on frequency data—as examined here—can provide valuable insight, there are potential concerns for the interpretation of the model terms in regression approaches of this type (Ai & Norton, 2003). For example, Ai and Norton (2003) demonstrate that interaction terms based on frequency data may occasionally show opposite sign, or may be sensitive to parameter effects. To address this concern and check for robustness of findings, the GLM results

on the frequency data can be compared with results from ordinary least squares (OLS) regression using the proportions (calculated as the frequency within each cell of the table divided by the total frequency in the table). To that end, each model specified in Figure 1 that uses a GLM on the frequency data is repeated with an OLS regression with the dependent variable as the proportions within each cell, and using the same independent variables. Note that, for the fully-saturated GLM models (i.e., models 1.1 and 2.1), the corresponding OLS regression would have 0 residual degrees of freedom—so no statistical tests can be conducted on these fully-saturated models. Even so, if findings from the frequency GLM are consistent with the OLS regression on proportions, this can lend weight to the findings by demonstrating robustness of the observed results.

Results

To compare findings from the GLM approach with the conventional alignment approach, results are presented for each data set, respectively. The New York Regents physics data are presented first, followed by the Virginia 6th-grade ELA data.

Data Set 1

Results from the GLM analyses for the first data set are shown in Table 2. Model 1.1 was the fully saturated model that is important as a comparison, but that cannot be analyzed further. Because it is fully saturated, it has 0 residual deviance and 0 residual degrees of freedom; it also has the highest AIC (215.37). Model 1.2 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 1.3 removes the joint dependence term for Cognitive Level and Topic and focuses only on dependence of Source with Cognitive Level

and with Topic. Finally, Model 1.4 is the fully independent model, without any joint dependence terms.

Model 1.3 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 between Models 1.3 and 1.4 (χ^2=19.35, df = 9, p<.05). Thus, Model 1.3—with joint dependence of Source with Topic and Source with Cognitive Level—is preferred as having superior model-data fit, so its terms are interpreted to understand the significant effects.

Under Model 1.3, there are significant main effects for both Cognitive Level and Topic, but non-significant effect for Source document. The main effect for Cognitive Level indicates that there are differences in the distribution of frequencies by the different cognitive demands. This makes sense, as cognitive demands such as recollection or understanding may be more frequent than create for both tests and standards. Similarly, the main effect for Topic indicates that both test and standard may emphasize a topic of the subject over others—such as having more questions on properties of matter than on waves.

There is also statistically significant interaction between Source and Cognitive Level. That is, the test and the standards have significantly different proportion of points by Cognitive Level. Examining the marginal discrepancies between the test and curriculum just for cognitive level (Figure 2), shows that this significant interaction exists because the test underemphasizes the skills of *understand* and *analyze*, but overemphasizes *recollection* and *application*. By contrast, there is not a significant interaction term for Topic with Source, so any marginal discrepancies by topic area as seen in Figure 3 are not statistically significant.

These findings can also be compared with an OLS counterpart to the GLM, and with typical alignment index analysis. For the OLS regression for the proportions (rather than

frequencies), the findings were consistent with the GLM results. The model with superior fit was the OLS counterpart to model 1.2, the joint dependence model with Content and CogLevel. Just as with the GLM, the OLS model shows a statistically significant interaction of Source with CogLevel ($F [5,20] = 5.07$, $p < .01$). Likewise, there is not a significant interaction effect of Source with Content ($F [4,20] = 0.34$, $p > .80$). This parallel of the OLS results with the GLM results demonstrates robustness of the findings from the GLM, and provides further evidence for attention to the Source-CogLevel interaction effect.

For the alignment index analysis, the Porter alignment index for this is 0.80. This index is statistically significantly different from what alignment index could occur by chance (0.689; Fulmer, 2011), equivalent to a z-score of 2.56 ($p<.05$). That is, the test has higher alignment than could have occurred by chance. However, it would not be possible using the Porter alignment method to determine whether the apparent differences are statistically significant—as is the case for cognitive level (Figure 2)—or not—as is the case for topic (Figure 3). Thus, the GLM findings provide an important complement to the alignment index approach that can allow further interrogation of potential types of misalignment between the source documents.

Data Set 2

Results from the GLM analyses for the second data set (from Polikoff, et al., 2011) are shown in Table 3, with the same model comparison process used for data set 1. As with the GLM for data set 1, Model 2.1 was the fully saturated model so has 0 residual deviance and 0 residual df; it also has the highest AIC for this data set (1587.8). Model 2.3 has the lowest AIC of the estimated models (906.2). Furthermore, likelihood ratio tests for the models show that the increase in residual deviance was significant for models 2.3 and 2.4 ($\chi^2=255.47$, df = 66, $p<.001$).

Thus, Model 2.3 is preferred as having superior model-data fit. This shows that the superior model has joint dependence of Source with Topic and with Cognitive Level.

Similar to the New York Regents physics data and consistent with expectation, the Virginia 6th grade ELA data show significant main effects for both Cognitive Level and Topic, and a non-significant main effect for Source document. The main effects can be interpreted to mean that both test and standard emphasize some topic areas or some cognitive demands over others.

The Virginia SOL shows a statistically significant interaction between Source and Cognitive Level, indicating a significant difference in the distribution of frequencies by cognitive level between the test and standards. This can be examined by graphing the marginal discrepancies by cognitive level, as shown in Figure 4. From Figure 4, it can be seen that the Virginia sixth-grade ELA test differs from the respective standards by less than 5% overall, which has much smaller magnitude than the differences observed for New York Regents physics, but it is significant in this case because there are more degrees of freedom (and hence greater power to detect difference) due to the higher overall frequencies of test items and content standards that are coded.

For the topics, there are also significant differences between the test and the standards, with the marginal discrepancies shown in Figure 5. As Figure 5 shows, there is a difference of up to plus or minus 6% across topic areas with a great deal of variation in relative emphasis between the test and standards. This contrasts with New York Regents physics, which did not have any interaction of topic by source.

These findings can be compared with an OLS counterpart to the GLM, and with typical alignment index analysis. For the OLS regression for the proportions (rather than frequencies),

the findings were consistent with the GLM results. That is, the OLS counterpart to model 2.3 had the best data-model fit, and shows a statistically significant interaction of Source with CogLevel ($F [4,496] = 6.34$, $p < .001$) and of Source with Content ($F [62,496] = 1.97$, $p < .001$). The OLS findings provide further evidence of the robustness of the findings from the GLM analysis. For the alignment index analysis, Polikoff and colleagues (2011) reported an alignment index of 0.31. Based on a simulation on alignment indices (Fulmer, 2011), the observed alignment index of 0.31 is not significantly different from the index that could occur by chance (0.294), with an equivalent z-score of 0.54 ($p > .10$). Thus, the traditional alignment analysis indicates the test is not any more or less aligned than could have occurred by chance under the coding conditions. The GLM analysis cannot test this overall alignment effect (as the three-way interaction will always have zero residual degrees of freedom). However, the GLM results do provide information about statistically significant differences between the test and standards according to both topic and cognitive levels. This information cannot be gained from the traditional alignment analysis, thus providing valuable and complementary evidence on alignment.

Discussion and Conclusions

In the present study, an alternative to typical alignment analysis is demonstrated. The typical alignment analysis allows the determination of the extent of alignment and whether there is, overall, a high or low alignment. But it cannot allow deeper interpretation of the findings. By comparison, drawing upon the GLM results enables such interpretations. So, for data set 1 (New York Regents physics) there is a significant difference between the source documents by cognitive level, but there is not a significant difference between the source documents by topic.

For data set 2 (Virginia's 6th-grade Standards of Learning), there are significant differences between the test and standards for cognitive level and for topic.

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 (2002) alignment index approach. This approach does not necessarily replace prior work on estimates of alignment (e.g., Fulmer, 2011; Porter, 2002); rather, it provides an additional method to test whether observed differences in ratings are statistically significant and to provide more insight into the types of misalignment that might exist. The analysis of alignment indices allows consideration of overall alignment, much like a "big picture" consideration of alignment between any two documents. This is particularly useful for policy decisions about the match of an assessment instrument or teachers' instruction with a state's standards documents. However, analyses of alignment using the GLM could be very effective for determining the ways in which a test and standards are misaligned, which would be essential for ensuring the content validity and consequential validity of such tests in their use to evaluate students, teachers, and schools (Messick, 1995).

Limitations

While the current results suggest promising direction for approaches to the study of alignment using GLM, the present study also has limitations. First, while the GLM approach is demonstrated to be promising for alignment analyses, it is not necessarily a replacement for alignment index approaches. Index analyses are based on cell-by-cell agreements or discrepancies, using Porter's method, which are similar to an interaction term of Topic×CogLevel×Source in the GLM approach. Yet the GLM regression cannot test the significance of this three-way interaction term, as this would constitute the fully saturated

statistical model. That is, it is not possible to say which specific *cells* have statistically significantly higher or lower alignment between source documents. As such, GLM approaches can complement but not replace traditional alignment index approaches. Thus, there remains significant room for future work to address this limitation and to extend this proposed method further to address this limitation.

Second, the study uses as examples just two sample data sets from previous work. As examples for the method, this article does not go into great detail on how the proposed method influences the interpretation of findings or how this affects the conclusions from the previous studies. Therefore, any reanalysis and reinterpretation of prior published studies would require more intentional analysis and comparison.

Implications for Policy, Theory, and Practice

The present study has implications for policy, theory, and practice. In terms of policy, the use of the GLM model will be very informative on one hand about policy decisions based on such tests. That is, policymakers, educational leaders, or researchers may wish to understand *how* a test and its standards are misaligned and use this to determine what changes are appropriate in test design or in the use of the test for high-stakes purposes. For example, in the case of data set 1, it is clear that the significant discrepancy is due to differences in the cognitive level of the test tasks, but not topics assessed by the test. Therefore, improving the test alignment would require adaptations to the cognitive skills that students need to use to answer items, but no changes would necessarily be needed in the proportion of items by topic.

Regarding implications for research, the present method is extensible to consider possible effects of multiple raters or other variations that reflect differences in practice among coding schemes. While this is beyond the current scope of the current study, the approach used here

based on three-way contingency tables could be expanded to more complex designs. Therefore, the application of GLM regression analyses to alignment studies has potential to increase the quality and depth of discussion around observed extent of alignment and the possible forms of discrepancy and misalignment that can exist. Subsequent researchers must be aware of the potential for opposite direction for interaction effects in complex models for frequency data (e.g., Ai & Norton, 2003), and follow the present example for a robustness check using both generalized analyses on the frequencies as well as ordinary least-squares analyses on the proportion data. This will help increase confidence in the results and interpretations.

The study also has potential implications for practice at the classroom level. The information from GLM results could be informative for teachers, who could use the results to consider what topics or cognitive skills to emphasize when preparing their students for high-stakes tests. Continuing the example of data set 1, teachers could decide to emphasize analysis and application skills more than the standards would suggest, thus helping their students prepare for the emphasis of the test.

Conclusion

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. With the proposed method for analyzing alignment among source documents, the present study provides another step toward providing tools that

researchers, educators, and policy makers can use to compare and interpret alignment or misalignment.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*(1), 123-129.

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, 22*(3), 21-29. DOI: 10.1111/j.1745-3992.2003.tb00134.x

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple correlation: Regression analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361

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

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741-749.

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. (2012a). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis, 34*(3), 278-294.

Polikoff, M. S. (2012b). Instructional alignment under No Child Left Behind. *American Journal of Education, 118*(3), 341-368.

Polikoff, M. S., & Fulmer, G. W. (2013). Refining methods for estimating critical values for an alignment index. *Journal of Research on Educational Effectiveness, 6*(4), 380-395. doi: 10.1080/19345747.2012.755593

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.

Virginia Department of Education. (2012). *Virginia Standards of Learning*. Richmond, VA: Commonwealth of Virginia. Retrieved from http://www.doe.virginia.gov/testing/.

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

Cognitive Level	Content Topics				
	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 for New York Regents physics test and curriculum

Source	df	Model			
		1.1	1.2	1.3	1.4
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

Note. Model 1.1 is the fully saturated model including the three-way interaction term. Model 1.2 has all 2-way interactions terms. Model 1.3 has only two-way interactions involving Source. Model 1.4 has only main effects

** p<.01; *** p<.001

Table 3

GLM Analysis Results for Four Nested Models for Virginia 6th grade SOL test and standards

Source	df	Model			
		2.1	2.2	2.3	2.4
CogLevel	4	101.19 ***	101.19 ***	101.19 ***	101.19 ***
Content	62	326.40 ***	326.40 ***	326.40 ***	326.40 ***
Source	1	0.06	0.06	0.06	0.06
Source×CogLevel	24 8	48.01 ***	48.01 ***	48.01 ***	
Source×Content	4	176.33 ***	176.33 ***	176.33 ***	
CogLevel×Content	62	286.57 *	286.57 *		
CogLevel×Content×Source	24 8	54.92			
Model Residual Deviance		0	54.92	310.36	565.83
Model Residual df		0	248	496	562
AIC		1587.8	1146.7	906.16	1029.6

Note. Model 2.1 is the fully saturated model including the three-way interaction term. Model 2.2 has all 2-way interactions terms. Model 2.3 has only two-way interactions involving Source. Model 2.4 has only main effects

** p<.01; *** p<.001

Figures

Data set	Model name	Model type	Description of model terms
Data set 1: New York Regents	1.1	Fully saturated	Fully saturated model, including all terms up to 3-way interactions
	1.2	Joint dependence	All 2-way interaction terms
	1.3	Joint dependence	Only 2-way interaction terms including Source
	1.4	Mutual independence	Only main effects
Data set 2: Virginia SOL	2.1	Fully saturated	Fully saturated model – all terms including 3-way interactions
	2.2	Joint dependence	All 2-way interaction terms
	2.3	Joint dependence	Only 2-way interaction terms including Source
	2.4	Mutual independence	Only main effects

Figure 1. List of generalized linear models tested. In all cases, the dependent variable is the frequency of curriculum or test points that are assigned to each cell.

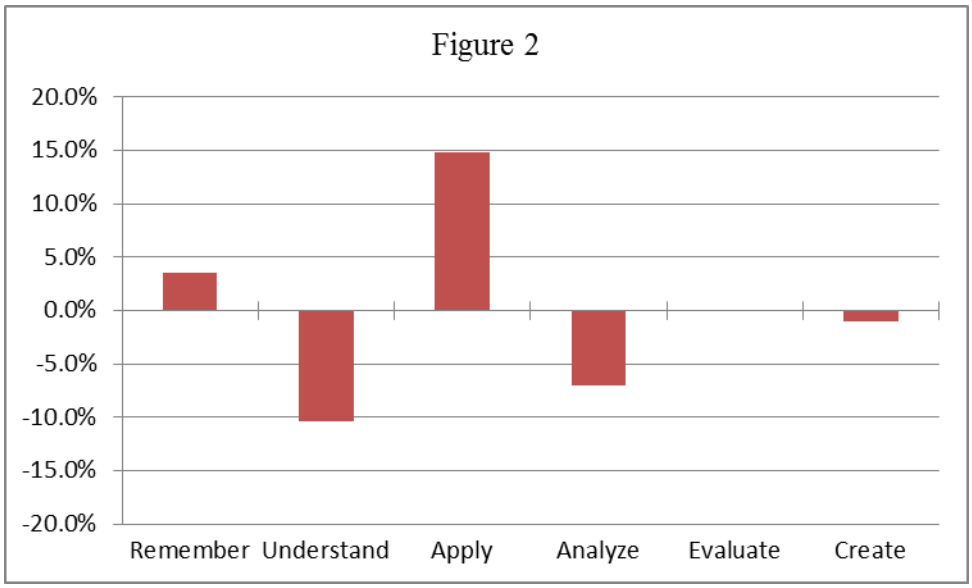

Figure 2. Chart showing marginal discrepancies for cognitive demands between the New York Regents test and curriculum. Positive values indicate the test shows greater emphasis than the standards on the respective cognitive level.

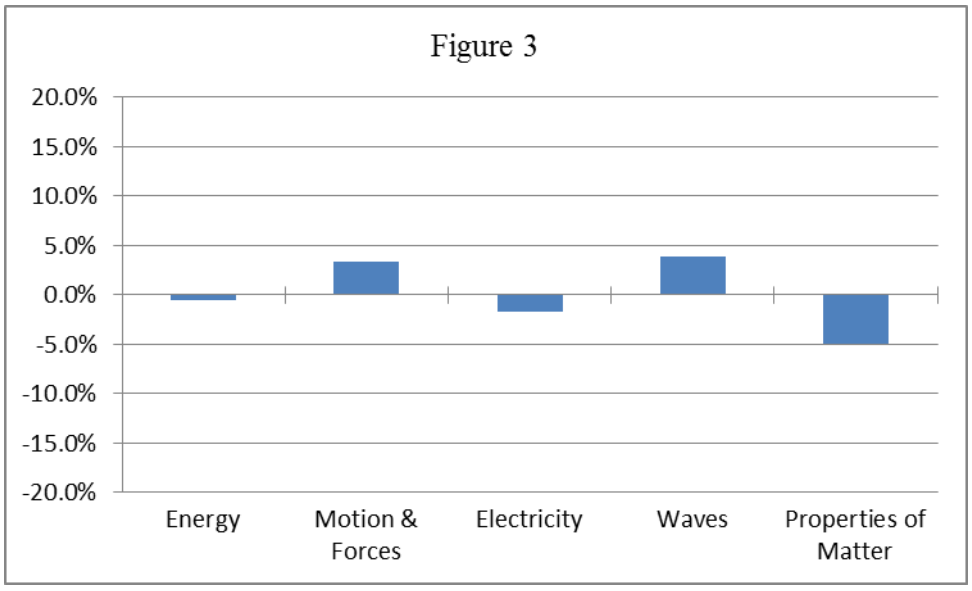

Figure 3. Chart showing marginal discrepancies for content areas between the New York Regents test and curriculum. Positive values indicate the test shows greater emphasis than the standards on the respective content area.

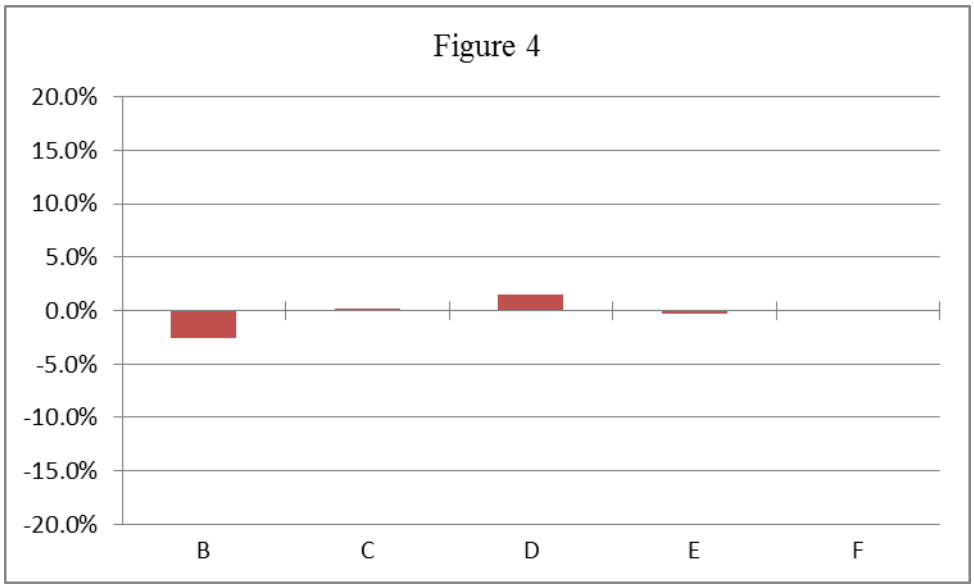

Figure 4. Chart showing marginal discrepancies for cognitive demands between the Virginia 6th-grade ELA test and standards. Positive values indicate the test shows greater emphasis than the standards on the respective cognitive level.

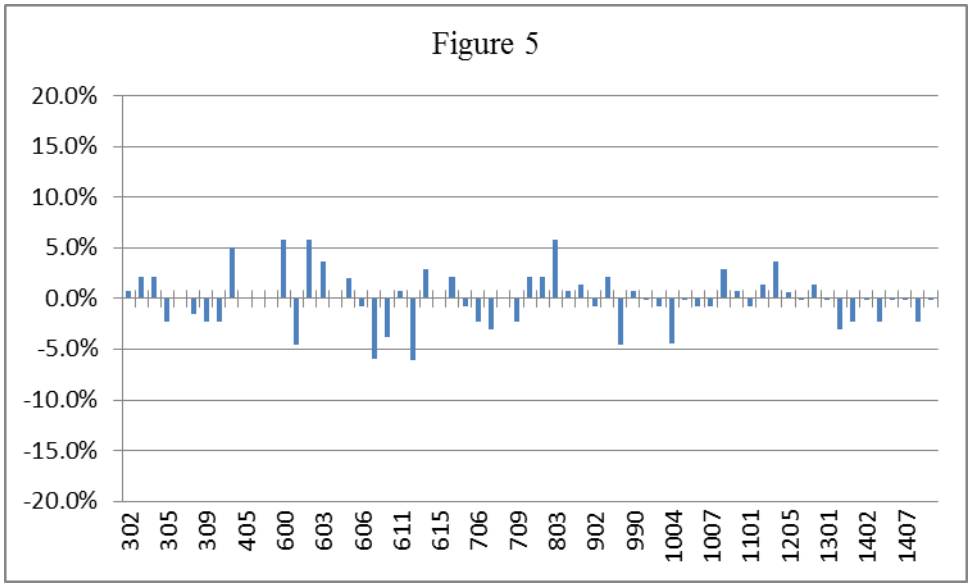

Figure 5. Chart showing marginal discrepancies for content areas between the Virginia 6th-grade ELA test and standards. Positive values indicate the test shows greater emphasis than the standards on the respective content area.