Educ Asse Eval Acc (2014) 26:177–201
DOI 10.1007/s11092-013-9186-6

Use of the probit model to estimate school performance in student attainment of achievement testing standards

W. Holmes Finch · Jerrell C. Cassady

Received: 27 September 2012 / Accepted: 30 December 2013 /
Published online: 18 January 2014

Abstract In the USA, trends in educational accountability have driven several models attempting to provide quality data for decision making at the national, state, and local levels, regarding the success of schools in meeting standards for competence. Statistical methods to generate data for such decisions have generally included (a) status models that examine simple indications of number of students meeting a criterion level of achievement, (b) growth models that explore change over the course of one or more years, and (c) value-added models that attempt to control for factors deemed relevant to student achievement patterns. This study examined a new strategy for student and school achievement modeling that augments the field through the use of the probit model to estimate the likelihood of students meeting an established level standard and estimating the proportion of individuals within a school meeting the standard. Results of the study showed that the probit model was an effective tool both for providing such adjustments, as well as for adjusting them based upon salient demographic variables. Implications of these results and suggestions for further use of the model are discussed.

Keywords School assessment · Standardized achievement test · Probit model

Perhaps the most pervasive and controversial topic in K-12 education has been the dramatic and accelerated increase in the use of standardized measures of student achievement to determine school effectiveness or quality. The authorization of the No Child Left Behind (NCLB) legislation and associated statewide school accountability laws, such as those mandating that all schools and students within schools participate in standardized achievement testing programs, has promoted an era of assessment that pins judgments regarding the performance of schools, teachers, and students to score patterns on statewide assessment models. While most in the field of education support strong

W. H. Finch (✉) · J. C. Cassady
Department of Educational Psychology, Ball State University, Muncie, IN 47306, USA
e-mail: whfinch@bsu.edu

measures of accountability to ensure that schools are achieving their mandates, there is considerable disagreement in the methods, materials, and judgment criteria used to make these high stakes decisions (Darling-Hammond et al. 2012). Thus, the primary goal of this study was to introduce a simple, but effective statistical tool to provide information to policy makers, teachers, parents, and other interested parties regarding the relative performance of students in different schools on large scale assessments, and to demonstrate how such measures can be made to account for the context in which the schools operate. Specifically, the method described here will make use of the probit regression model in conjunction with a dichotomous outcome variable indicating whether or not examinees met a particular academic standard on an achievement test. This context is different than what has been proposed in earlier models of academic achievement, such as the value-added model, which focus on the actual scores that examinees attain on such tests. However, in many educational contexts, academic performance is reported only in terms of whether an examinee has met a given standard or not. Thus, the current study adds to the literature by introducing a statistical modeling method that is more appropriate than previously used approaches in such dichotomous testing outcome situations.

1 Assessment of school effects

In the USA, a variety of strategies for determining if schools are meeting the performance criteria established by federal or state guidelines have been used over the past few decades. A quick review of these primary methods (i.e., status, growth, and value-added models) is offered to identify the methodological and philosophical stances that drive decision making in school accountability. Three primary themes for these assessment strategies are quite common. First, criterion-based models identify a standard of performance that is considered to demonstrate students' mastery of a core set of academic tasks (e.g., math, reading, language arts). Naturally, students who reach or exceed the criterion point are deemed to "pass," and school effectiveness is essentially determined based on the percentage of students who achieve this distinction. The consequences of failing to have sufficient percentages of students reach the criterion performance level are dire—schools face remediation plans, restructuring, or closure over a period of years of continuously failing to meet the mark (Linn 2000). Critics of this approach hold that the determination of school quality in this model for evaluation does not account for the myriad of contributing factors that dictate eventual student success on academic measures (Downey et al. 2008). Factors such as the socioeconomic status, race, geographic region, and pre-K educational opportunities of students attending the schools have all been identified as key variables that influence the probability of a student achieving established standards of performance on standardized tests (Baker and Johnston 2010; Capraro et al. 2009; De Lisle et al. 2010; Perry and McConney 2010; Wiggan 2007). This raises the common argument that accountability systems are biased, unfair, or at the minimum provide an incomplete method for determining the overall quality of educational instruction provided at individual schools (Goldschmidt et al. 2005).

2 Measuring student growth

Many districts and states have augmented this standard criterion (or “status”) model for determining school effectiveness by examining measures of growth within schools (Scherrer 2012). Early approaches focusing on academic achievement growth were often unreliably simplified—examining group trends over time within or between cohorts to determine if schools were demonstrating gains in the percentage of individuals reaching the established cut scores for each grade level assessment (Barton 2008; Lee 2010). Particularly in the USA, more sophisticated analyses have been developed and refined in recent years by statisticians for use in large scale, often statewide testing programs, with the implementation of individual student growth analyses that compare achievement growth over time relative to regional or statewide norming samples (Chiang 2009).

3 Limitations of student growth indicators

While these growth indicators provide an advancement in determining the impact of schools (or teachers) on student achievement, they remain largely uninfluential in the broad school accountability discussion for a variety of reasons. First, while some policy guidelines applaud the benefits of examining growth, the reality is that growth measures will not meet the requirements of meeting NCLB legislation alone—requiring the status indicators as the primary value of importance (Goldschmidt et al. 2005). As such, those schools with low initial performance may demonstrate growth over time, but still fall short of the school wide success indicators because the requisite level of growth is unattained. Second, as with basic criterion or status accountability models, the probability of student achievement growth over time is also influenced by factors external to the schools (Downey et al. 2008). In reality, students from at risk populations generally fall further and further behind their classmates over time. As such, they not only fail to reach the criterion mark for success, but they also show lower relative growth over time. Lastly, the accuracy of growth models is unreliable at times by overestimating the number of students who will reach proficiency (Weiss and May 2012). The unreliable nature of growth models is also dependent on the type of scale used. For example, normal curve equivalents are particularly likely to underestimate individual student growth, and standards-based testing is impacted by school size (Goldschmidt et al. 2010).

4 Value-added models

A third strategy for assessing school and teacher performance, in terms of their efforts to provide quality instruction to students, that has gained some traction in the accountability literature has focused on examining the influence of various factors at the school or individual level to determine the impact of the students’ educational experiences on their school achievement and growth (Darling-Hammond et al. 2012). This focus on examining influences of multiple factors on student performance has led to changes in state assessment practices, particularly with regard to how achievement testing data are analyzed and reported to schools and districts, regarding how schools are rated and ranked based on student achievement, and in some cases how resources are allocated to schools and school districts (McCaffrey 2013). In general, these strategies examine the performance of the student while taking into consideration the contexts within their education has taken place, examining regional, school, teacher, and individual differences for impact (Braun 2005). Proponents of these strategies propose that this methodological approach provides greater precision in identifying relative quality of education by statistically controlling for factors that are known to influence performance patterns of school-age children (Scherrer 2012) and thereby providing estimates of academic performance that are adjusted for such contextual factors (e.g., receipt of free or reduced lunch). One broad class of analyses that are dominant in these discussions is the value-added modeling approach (VAM). VAM is worthy of some discussion here both because it is becoming more widely used in conjunction with state testing programs across the USA and because it has been shown to effectively account for contextual factors that are related to achievement score growth, such as socioeconomic status (McCaffrey et al. 2004). The purpose of VAM in school accountability is to essentially control for the known external variables that influence performance (e.g., family and community factors related to achievement) and examine the rate of growth that the child demonstrated over a period of time in a given educational context as compared to the predicted growth based on those factors.

The results obtained from VAM are generally offered as indicators of the “effectiveness” of a given teacher or school—those schools or teachers whose students perform at levels higher than the predicted level of growth (or basic achievement) are seen as more effective than those who do not demonstrate a significant “value-added impact.” In theory, this model of assessing school or teacher effects has merit. Attempting to control for prior achievement, family and community variables, and regional effects appears to place teachers and schools on an equal footing (Scherrer 2012). However, the results have demonstrated that there is low consistency across years for both schools (Gorard 2011) and teachers (Darling-Hammond et al. 2012). That is, when examining the VAM for school or teacher effects from year to year, there is a high degree of “flipping” in determining the effects—where highly effective schools and teachers are unlikely to be consistently identified as such over time. If the impact of the teacher or school is accurately evaluated by VAM, it would be reasonable to expect a more consistent trend over time—as teacher quality can be assumed to relatively stable and the VAM approach takes into consideration the individual variances associated with student effects. Additionally, VAM requires vertical scaling, so that if the scaling method (i.e., IRT, calibration, and student proficiency estimation) differs from year to year, the VAM can be adversely impacted (Briggs and Weeks 2009). Such sensitivity can lead to inaccurate classification of schools in terms of their success or failure.

5 Importance of contextual variables in assessing school performance

As described above, education policy makers and teachers are frequently interested in whether students are able to meet certain performance standards on standardized achievement measures. Very often, these outcomes are expressed as meeting the standard or not, based upon performance on an achievement test, and results are expressed as simple proportions meeting and not meeting the standards (Goldschmidt et al. 2005). A potential problem with such simple reporting practices is that they do not take into account school context variables such as distribution of students by grade or demographic factors. Indeed, there is not clear agreement on the optimal approach for determining when schools have, and have not, demonstrated sufficient academic performance (Darling-Hammond et al. 2012; Franco and Seidel 2012). The VAM approach has become very popular in large part because it allows for assessment of school or teacher performance, while accounting for the impact of external variables on student achievement (Olson 2004). However, as also previously discussed, VAM can be very sensitive to a myriad of effects, leading to inconsistent results over time (e.g., Lockwood et al. 2006). In addition, the inclusion of external variables into the model has been shown to raise questions of whether, and to what extent, the construct of interest continues to be accurately modeled and appropriate weights applied in the context of VAM (Martineau 2006; Schmidt et al. 2005). The manner in which such variables should be included is also an issue of some discussion, particularly with regard to whether they should be used as covariates in a statistical model, or for matching individuals using propensity scores (Ballou et al. 2004). Finally, VAM is typically applied to scores, rather than individual markers of performance such as pass/non-pass. However, in many large-scale state testing programs, student performance is reported in a dichotomous manner, and school level results are most often presented in terms of the proportion of students meeting the standard (Goldschmidt et al. 2005, 2010).

Given the dichotomous nature in which student performance is frequently reported in large-scale testing programs (e.g., met the standard or not), we would argue that a useful modeling approach that takes into account this dichotomy should meet three criteria: (1) It uses the most common metric for reporting such results, the proportion meeting a standard; (2) It allows for the adjustment of results by pertinent external variables in order to provide a clearer picture of relative school performance; and (3) It is relatively simple so as not to require a number of statistical assumptions about the data, violation of which could lead to inconsistent results such as those evident with VAM. The remainder of this manuscript proposes the use of the probit regression model for this purpose, as it meets these three criteria and provides potentially useful information regarding relative school performance.

6 Probit model

Researchers interested in investigating a dichotomous dependent variable, such as whether examinees meet an achievement test growth standard for example, would typically use one of two statistical approaches, logistic or probit regression. While both models are designed for dichotomous outcomes, they differ in terms of their assumptions about the underlying process that created the dichotomous variable, leading to a difference in the type of dependent variable transformation (link function) that is used to create a linear relationship with one or more independent variables. Logistic regression involves the logit link, which takes the form

$$ln\left(\frac{\pi}{1-\pi}\right). \tag{1}$$

Here, π is the probability that the event of interest occurs, e.g., the examinee meets the achievement test standard. The logit can be interpreted as the log of the odds that the event will occur, e.g., the student will meet the performance standard. For a predictor variable, x, that we believe is related to this binary outcome, the logistic regression model can be expressed as

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x, \tag{2}$$

where α is the intercept and β is the slope relating x to the outcome. The slope is of particular interest in this model, as it expresses the nature and strength of the relationship between x and the log of the odds that the outcome of interest will occur. Choi and Goldschmidt (2012) used logistic regression to analyze a 3-year set of growth data for the California High School Exit Exam, demonstrating the efficacy for predicting success in passing the final outcome exam by using prior performance measures and, to a lesser degree, the growth trends over 3 years of assessment. Their findings highlighted the potential utility of using dichotomous variable models in assessing student performance. However, their primary goal was not that of the current study, namely to estimate school performance in terms of proportion of students meeting an academic standard. Rather, logistic regression was used to predict individual student performance and identify students at potential risk for failing the high school exit exam.

Although logistic regression is perhaps the most commonly used approach for modeling a binary outcome, another link function that is also appropriate for such data is the probit. Given the probability of the outcome of interest π, the probit regression model is written as

$$\Phi^{-1}(\pi) = \alpha + \beta x. \tag{3}$$

The probit link function is the inverse of the cumulative normal distribution function and essentially converts the probability of the outcome of interest occurring (e.g., student meeting the growth standard) to the z score corresponding to that probability in the standard normal distribution. This link transformation assumes that the process underlying the probability distribution is continuous and in fact normally distributed. In other words, if the outcome of interest is whether an examinee meets a particular performance standard on an achievement test, use of the probit model implies a belief that the actual outcome of interest (e.g., academic performance) lies on a continuum, and π is the probability that an individual's performance is likely to exceed a threshold so that the observed outcome is pass the test, in this case. For example, an individual, i, with a π_i of 0.9 is quite likely to meet the academic standard, i.e., likely to surpass the threshold on the underlying continuous variable of student achievement. Conversely, someone with a π_i of 0.2 is highly unlikely to pass the threshold and thus be deemed to have mastered the tested material.

7 Probit regression as a means for understanding school effects

While both the logistic and probit models are appropriate for use with binary outcome variables, they assume different underlying mechanisms about the process under study.

The logistic regression model posits that the outcome is a true dichotomy, so that all individuals with the same value are homogeneous. Conversely, as noted above, the probit model assumes that the actual outcome of interest is a normally distributed random variable that is manifested as a dichotomy based upon whether an individual has surpassed the threshold (Azen and Walker 2011). Thus, even while they may provide similar results with respect to model parameter estimates such as the slope, they are undergirded by very different assumptions about the actual data-generating mechanism. In the context of educational testing, the underlying data is indeed quite often a continuous variable, in the form of a test score measuring some aspect of academic achievement. Examinees whose scores exceed a predetermined cut value are given a passing outcome, while those with scores below the cutoff are not. Given this common approach to assessing student performance, the probit model appears to be ideally suited to studying achievement test outcomes coded as either meeting or not meeting a standard, when that decision is based upon performance on a standardized test.

In addition to providing information regarding relationships between potentially salient independent variables and student achievement measured in a dichotomous fashion, the probit model can also be used to obtain estimates of the proportion of individuals meeting the standard, adjusted for these independent variables (Agresti 2002). For example, consider the probit model in which the outcome of interest is whether an examinee met the testing standard (yes or no), and the sole independent variable is the school attended by each student. Using this model, it is possible to obtain an estimated mean for each school, which is simply the proportion of examinees at the school who met the standard. If an additional variable is included in the model, for example grade level, a mean estimate can once again be obtained for each school. However, for this second model, the mean will be adjusted to reflect the impact of this additional variable on the likelihood of the examinee meeting the standard. The adjusted mean is calculated as

$$\overline{y}_{aj} = \overline{y}_j - b\left(\overline{x}_j - \overline{x}\right), \tag{4}$$

where $\overline{y}_j$ is the unadjusted mean for group j (e.g., the proportion meeting the standard for school j), b is the slope relating the outcome to the additional covariate (e.g., grade), $\overline{x}_j$ is the mean of the covariate for group j (e.g., the mean grade for school j), and $\overline{x}$ is the overall mean on the covariate (e.g., mean grade across schools). Thus, this adjusted mean will reflect the school proportion meeting the standard when differences on the distribution of grade by school are accounted for. Schools with covariate means further from that of the sample as a whole will experience a greater adjustment in the outcome mean than those schools with covariate means close to the overall mean. For the probit model with a dichotomous outcome, these adjusted means essentially reflect the proportion of students meeting the standard after the impact of the salient independent variables has been accounted for. These adjusted proportions can, in turn, provide potentially more accurate information regarding the relative performance of individual schools, while appropriately accounting for the underlying continuum of achievement test scores upon which the classification regarding standard meeting by individual students was made.

8 Goal of the current study

As discussed above, one of the most common examples of using statistical models to examine school effects is the VAM. VAM has proven popular because of its ability to assess the effects of teachers and schools, after accounting for the impact of other salient variables particularly at the student level (Braun 2005). However, its use is typically restricted to continuous outcome variables such as scores on an achievement test, or change in such scores, rather than situations in which the meeting of a standard is of primary interest. In some contexts, though, policy makers are primarily interested in whether examinees meet a particular standard and in whether there is a relationship between the school attended and the meeting of the standard. In such cases, the VAM will not prove useful. On the other hand, simple reporting of proportions do not reveal (1) if students at specific schools are significantly more or less likely to meet the standard than students in general, and (2) how estimates of the proportion meeting the standard might change when additional information about the students that is salient to their performance is accounted for. Therefore, the goal of this study was to demonstrate the utility of the probit model for estimating the impact of individual schools on the likelihood of an examinee meeting a particular test standard and to demonstrate how relative school performance changes when pertinent demographic information about students is considered.

9 Methodology

9.1 Subjects

Data were drawn from 8,779 elementary school students (grades 1–8) attending 41 charter schools in a Midwestern state. Table 1 includes demographic information for the study participants. Approximately half of the students included in the study were female, and the majority was non-Caucasian. In addition, relatively few students received special education or Title 1 services or were English as a second language (ESL) learners. Table 2 includes information on the grade levels of participants. Among these elementary students, representation was approximately equivalent across grades, with slightly lower representation among eighth graders when compared with the others. The charter schools included in the sample were selected for inclusion in this study for two reasons. First, these schools had available testing data from the Northwest Evaluation Association (NWEA) assessment, which has been shown to be of especially high quality, particularly with respect to measuring academic growth (Capraro et al.

Table 1 Demographic information for study participants: percent (*N*)

Variable	*N*=8,779
Female	49.4 % (4,337)
Title 1	0 % (0)
Special education	2.2 % (193)
English as a second language	0.9 % (79)
Caucasian	22.8 % (2,002)

Table 2 Grade level by type of school

N=8,779	
Grade	Percent (*N*)
1	12 % (1,053)
2	13.3 % (1,168)
3	13.6 % (1,194)
4	13.5 % (1,185)
5	13.8 % (1,212)
6	12.5 % (1,097)
7	11.6 % (1,018)
8	9.7 % (852)

2009). Given that such growth is of key importance in many academic settings, as well as for the current study, these data were seen as particularly advantageous for this work. The second reason that these schools were selected is that charter schools represent a particularly fast growing segment of the educational community (Aud et al. 2013). Thus, we felt that using data from a set of charter schools would be interesting and useful to researchers and policy makers alike. It should be noted that the method for assessing school performance that we demonstrate here is not limited to use with such schools, but can clearly be applied in any academic setting where educators are interested in assessing school and student performance with regard to achieving a specific academic goal, such as meeting a particular growth standard, or achieving a score at a particular level. At the same time, we recognize that charter schools do represent a unique population and that further research with this method should be extended to other types of public schools, as well as to private institutions.

9.2 Measures

The outcome measure for this study was whether examinees met the expected level of growth in the Measures of Academic Progress (MAP) reading, mathematics, and language assessments published by the NWEA between the Fall 2010 and Spring 2011 test administrations. MAP is a computer adaptive test (CAT) that selects the items given to individual students based on their ability level. In general, CATs administer items to students based upon a combination of item difficulty and student ability, as estimated using their responses to previous items. At the beginning of the test, all examinees are administered an initial item that is of moderate difficulty. Those who respond correctly are next given a slightly more difficult item, while those who respond incorrectly are given an easier item. This pattern of item administration continues, whereby examinees incorrectly answering an item in the sequence are given an easier item until they provide a correct response, while those answering items correctly receive increasingly difficult items until they respond incorrectly. When the statistical algorithm underlying MAP is able to converge on an estimate of an examinee's ability, the test is concluded. Using such a testing algorithm, CATs are better able to pinpoint the actual ability level of examinees on the construct being assessed (Wainer 2000). The MAP tests yield scores that are covalent, meaning that they can be compared interchangeably across grade levels, making the direct measurement of academic growth possible (Northwest Evaluation Association 2003). Actual test scores are expressed using a standardized metric based upon Item Response Theory, called the “Rasch Unit,” or RIT score, which ranges from approximately 150 to 300.

The growth standard to be met is equal to the median growth in the national norming sample for individuals at a particular RIT score from the fall test administration. For example, the expected growth standard for examinees with a Fall 2010 reading RIT score of 200 is equal to the median change in scores from fall to spring of all individuals in the norming sample who had an RIT value of 200 in the fall semester when they took the reading test. The change in RIT scores from Fall 2010 to Spring 2011 for each member of the current sample was compared to these median growth values from the norming sample. If an examinee’s actual growth equaled or exceeded this median norm sample growth, they were given a value of 1 for the outcome variable, while if their growth were less than the median, they were coded as 0. A separate such designation was made for the reading, mathematics, and language tests. Other variables included in the analysis were school attended, grade level, gender, and ethnicity (Caucasian or not). Several variables of potential interest that were not included in the analysis were special education status, Title 1 status, ESL, and free/reduced lunch. The first three of these variables were excluded due to the extremely small numbers represented in the sample, as can be seen in Table 1. Some schools had no such students, creating problems for model convergence when they were included in the analysis. In the case of free/reduced lunch, the opposite problem was present, as some schools had essentially all students in such a program, again creating intractable analytic problems with regards to model convergence. Thus, the exclusion of these variables from the analyses was due to practical issues rather than substantive beliefs about their relative importance in terms of student achievement. Indeed, were the distributions more favorable in terms of sample sizes, they would have been included in the models studied here.

9.3 Data analysis

As stated above, the goal of this study was to model the probability of examinees meeting a growth target as a function of school, as a way of characterizing typical student performance for a given school. Data analysis techniques, such as ordinary least squares regression, which are designed for continuous variables are not appropriate for use with dichotomous outcomes such as whether an individual met a particular achievement goal or not. In these cases, the assumptions of normality and homogeneity of variance are nearly always violated (Fox 2008). However, models for dichotomous variables do exist, including both logistic and probit regression. Statistically speaking, these models differ in that logistic regression is based upon the logistic cumulative density function, whereas the probit model is based upon the inverse normal distribution. In terms of application, researchers often make use of the probit model when the data-generating process underlying the outcome variable is known, or believed, to be continuous in nature (Agresti 2002). On the other hand, when the data-generating process is truly dichotomous, logistic regression is the method of primary interest (Agresti). Given that the data-generating process for the outcome in the current study was continuous in nature (change in score over time), the probit model was used, with school being a random effect in the analysis, and whether an examinee met the growth standard (yes/no) being the dependent variable. More specifically, the dependent variables in the probit analyses were as follows: student met the standard in reading (yes or no), student met the standard in math (yes or no), and student met the standard in language (yes or no). The proportions of examinees meeting the growth standard for each variable by school were then estimated based on this model. In order to identify schools that had exceptionally high or low rates of students meeting the standard, the deviance contrast comparing each school proportion to the overall statewide proportion was used in order to identify which schools had students performing significantly better, statistically the same as, or significantly worse than the set of schools as a whole. In order to control for type I error inflation due to conducting several such contrasts, Sidak's method was used (Tong and Lim 1980; Sidak 1967).

Three probit models were estimated for each achievement test (reading, math, and language), in order to control for and examine the impact of additional variables that have been shown to be related to student achievement. The first model included only school, model 2 included school and grade level, and model 3 included school, grade level, gender, and ethnicity. For each model, the proportion of individuals meeting the standard was estimated for each school, and the deviance test comparing the school proportions to the statewide proportion, across schools, was conducted as well. For the models including school only, the proportion was simply the sample proportion of examinees meeting the growth standard. In the other models, this proportion was adjusted for the inclusion of variables believed to be pertinent to the outcome. Of primary interest in this study was the examination of these school proportions, and the corresponding contrast results, and how these did or did not change across models. All analyses were conducted using the IBM SPSS Statistics software, version 19 (IBM Corp. 2010).

10 Results

10.1 Reading

Table 3 includes significance test results for each model. The significant results for school indicate that across models the schools had significantly different proportions of students meeting the growth standard on the reading exam. Furthermore, including other variables such as grade (model 2) or grade, gender, and ethnicity (model 3) did not change the fact that there were significant differences in the proportion of schools meeting the reading growth standard. In addition, grade was significantly related to the outcome, with the negative slope estimate (−0.012) indicating that examinees in higher grades were less likely to meet the standard than those in lower grades. Neither ethnicity nor gender was significantly related to the likelihood of an individual meeting the expected growth standard for reading.

Table 4 includes the estimated proportion of students meeting the reading standard for each of the three models at the state level and by school. Across schools, approximately 47 % of examinees met the expected growth standard in reading obtained from the median of the national norming sample, when only school was considered. With regard to the individual schools, the rate ranged from 32 to 73 %. Table 5 contains the

Table 3 Significance test results for reading exam, independent variable and school level

Variable	Model 1	Model 2	Model 3
School	Chi-square=394.171 *df*=43 *p*<0.001	Chi-square=385.353 *df*=43 *p*<0.001	Chi-square=360.523 *df*=43 *p*<0.001
Grade		Chi-square=5.069 *df*=1 *p*=0.024 *b*=−0.012	Chi-square=4.978 *df*=1 *p*=0.026 *b*=−0.012
Gender			Chi-square=0.011 *df*=1 *p*=0.916 *b*=−0.003
Caucasian			Chi-square=2.148 *df*=1 *p*=0.143 *b*=−0.062

results of the deviance contrast tests for each school. As a reminder, this test was used to compare the proportion passing in each school with the statewide proportion of 0.47. Sidak's method was used to control the type I error rate. Based on these results, schools 3, 6, 11, 28, 29, 31, 34, 38, and 40 all had significantly higher rates of examinees meeting the standard than the state average. On the other hand, schools 13, 30, and 33 had significantly lower rates of examinees meeting the standard when only school was considered.

Model 2 included both the school and grade level, accounting for the grade in which the student was enrolled. The purpose of this part of the analysis was to ascertain the extent to which inclusion in the statistical model of additional factors strongly believed to be related to the likelihood of a student meeting the academic growth standard would alter the estimated performance of individual schools. In this case, school officials were aware that the likelihood of an examinee meeting the academic standard was not consistent across grades. Given this fact, schools could be advantaged or disadvantaged in terms of their relative performance depending upon the mix of grade levels that they served. Therefore, in order to ensure that estimates of school effects were comparable regardless of their grade mixtures, we wanted to account for grade in the probit model and thereby obtain more accurate comparative school effects. As noted above, results in Table 3 reveal a significant negative relationship between grade and the likelihood of an examinee meeting the growth standard, so that children in higher grades were less likely to meet the standard. An examination of Tables 4 and 5 shows that results for the majority of schools did not change very much with regard to the reading outcome variable. However, school 20 had a significantly lower proportion of examinees meeting the growth standard when both school and grade were accounted for, which was not the case when only school was included in the model. This result suggests that once the impact of grade was removed from the estimate of the school effect, school 20 performed significantly worse than did examinees in the state as a whole. On the other hand, when grade was accounted for, school 30 was no longer shown to have

Table 4 Proportion of examinees meeting the growth standard for elementary reading test

School (*N*)	School only	School and grade	School, grade, sex, race
State average	0.47	0.47	0.48
1 (247)	0.59	0.59	0.60
2 (203)	0.52	0.54	0.54
3 (576)	0.73	0.73	0.74
4 (535)	0.53	0.53	0.54
5 (176)	0.52	0.53	0.58
6 (416)	0.69	0.68	0.69
7 (215)	0.59	0.59	0.61
8 (258)	0.49	0.49	0.46
9 (163)	0.53	0.53	0.54
10 (88)	0.64	0.63	0.64
11 (380)	0.65	0.65	0.66
12 (363)	0.47	0.47	0.50
13 (114)	0.36	0.36	0.37
14 (201)	0.49	0.48	0.49
15 (486)	0.56	0.56	0.57
16 (80)	0.47	0.47	0.48
17 (205)	0.43	0.44	0.46
18 (223)	0.51	0.51	0.52
19 (60)	0.53	0.53	0.58
20 (530)	0.45	0.45	0.46
21 (412)	0.46	0.46	0.47
22 (691)	0.52	0.51	0.52
23 (359)	0.53	0.52	0.52
24 (242)	0.44	0.44	0.44
25 (84)	0.46	0.45	0.45
26 (423)	0.55	0.55	0.56
27 (40)	0.36	0.38	0.36
28 (481)	0.61	0.60	0.61
29 (178)	0.71	0.72	0.73
30 (218)	0.40	0.42	0.39
31 (409)	0.64	0.63	0.64
32 (88)	0.61	0.61	0.61
33 (156)	0.32	0.32	0.34
34 (116)	0.66	0.67	0.67
35 (175)	0.59	0.59	0.60
36 (104)	0.57	0.57	0.57
37 (881)	0.50	0.51	0.52
38 (116)	0.68	0.68	0.69
39 (124)	0.65	0.65	0.66
40 (449)	0.64	0.64	0.65
41 (217)	0.54	0.53	0.55

Table 5 Deviance contrast results for elementary school reading test: (difference, *p* value)

School	School only	School and grade	School, grade, sex, race
1 (247)	0.08 (*p*=0.544)	0.07 (*p*=0.616)	0.07 (*p*=0.640)
2 (203)	0.01 (*p*=1.000)	0.02 (*p*=1.000)	0.01 (*p*=1.000)
3 (576)	0.21 (*p*=0.000)	0.21 (*p*=0.000)	0.21 (*p*=0.000)
4 (535)	0.02 (*p*=1.000)	0.01 (*p*=1.000)	0.01 (*p*=1.000)
5 (176)	0.00 (*p*=1.000)	0.01 (*p*=1.000)	0.05 (*p*=1.000)
6 (416)	0.17 (*p*=0.000)	0.16 (*p*=0.000)	0.16 (*p*=0.000)
7 (215)	0.07 (*p*=0.732)	0.08 (*p*=0.649)	0.08 (*p*=0.471)
8 (258)	−0.03 (*p*=1.000)	−0.03 (*p*=1.000)	−0.07 (*p*=1.000)
9 (163)	0.01 (*p*=1.000)	0.01 (*p*=1.000)	0.01 (*p*=1.000)
10 (88)	0.12 (*p*=0.570)	0.12 (*p*=0.639)	0.11 (*p*=0.715)
11 (380)	0.13 (*p*=0.000)	0.13 (*p*=0.000)	0.13 (*p*=0.000)
12 (363)	−0.05 (*p*=0.938)	−0.05 (*p*=0.908)	−0.03 (*p*=1.000)
13 (114)	−0.16 (*p*=0.019)	−0.16 (*p*=0.016)	−0.16 (*p*=0.013)
14 (201)	−0.03 (*p*=1.000)	−0.04 (*p*=1.000)	−0.04 (*p*=1.000)
15 (486)	0.04 (*p*=0.975)	0.04 (*p*=0.981)	0.04 (*p*=0.986)
16 (80)	−0.05 (*p*=1.000)	−0.05 (*p*=1.000)	−0.05 (*p*=1.000)
17 (205)	−0.09 (*p*=0.328)	−0.08 (*p*=0.713)	−0.07 (*p*=0.877)
18 (223)	−0.01 (*p*=1.000)	−0.01 (*p*=1.000)	−0.01 (*p*=1.000)
19 (60)	0.01 (*p*=1.000)	0.02 (*p*=1.000)	0.05 (*p*=1.000)
20 (530)	−0.07 (*p*=0.100)	−0.07 (*p*=0.041)	−0.07 (*p*=0.065)
21 (412)	−0.06 (*p*=0.578)	−0.06 (*p*=0.421)	−0.06 (*p*=0.515)
22 (691)	0.00 (*p*=1.000)	0.00 (*p*=1.000)	−0.01 (*p*=1.000)
23 (359)	0.01 (*p*=1.000)	0.00 (*p*=1.000)	−0.01 (*p*=1.000)
24 (242)	−0.08 (*p*=0.420)	−0.08 (*p*=0.535)	−0.09 (*p*=0.330)
25 (84)	−0.05 (*p*=1.000)	−0.07 (*p*=1.000)	−0.08 (*p*=0.999)
26 (423)	0.03 (*p*=1.000)	0.03 (*p*=1.000)	0.03 (*p*=1.000)
27 (40)	−0.15 (*p*=0.505)	−0.14 (*p*=0.721)	−0.17 (*p*=0.832)
28 (481)	0.09 (*p*=0.002)	0.09 (*p*=0.003)	0.08 (*p*=0.012)
29 (178)	0.19 (*p*=0.000)	0.20 (*p*=0.000)	0.20 (*p*=0.000)
30 (218)	−0.11 (*p*=0.018)	−0.10 (*p*=0.068)	−0.14 (*p*=0.257)
31 (409)	0.12 (*p*=0.000)	0.11 (*p*=0.000)	0.11 (*p*=0.000)
32 (88)	0.09 (*p*=0.976)	0.09 (*p*=0.963)	0.08 (*p*=0.991)
33 (156)	−0.20 (*p*=0.000)	−0.20 (*p*=0.000)	−0.19 (*p*=0.000)
34 (116)	0.15 (*p*=0.018)	0.15 (*p*=0.017)	0.14 (*p*=0.030)
35 (175)	0.07 (*p*=0.705)	0.07 (*p*=0.728)	0.07 (*p*=0.849)
36 (104)	0.05 (*p*=1.000)	0.05 (*p*=1.000)	0.04 (*p*=1.000)
37 (881)	−0.01 (*p*=1.000)	−0.01 (*p*=1.000)	−0.01 (*p*=1.000)
38 (116)	0.16 (*p*=0.011)	0.16 (*p*=0.009)	0.16 (*p*=0.008)
39 (124)	0.14 (*p*=0.059)	0.13 (*p*=0.071)	0.13 (*p*=0.075)
40 (449)	0.12 (*p*=0.000)	0.12 (*p*=0.000)	0.12 (*p*=0.000)
41 (217)	0.02 (*p*=1.000)	0.01 (*p*=1.000)	0.02 (*p*=1.000)

significantly lower performance than the state as a whole. An examination of the grade distributions for these two institutions revealed that in school 20, over 41 % of the students were in grades 1 and 2, compared with just 25.3 % in the state as a whole. Therefore, when this relative advantage of having more early grade students (who tend to perform better) was removed, the school's performance relative to the state as a whole declined. In contrast, 79.3 % of students attending school 30 were in grades 6, 7, and 8, which were the poorest performing grades across the state. Thus, when this relative disadvantage of having more older children than is typical was accounted for, school 30 performed comparable to the state as a whole, rather than worse.

Finally, model 3 included school, grade, gender, and whether or not the individual was Caucasian. As noted previously, there was not a significant relationship between meeting the standard and the gender or ethnicity of the examinee. This general lack of an effect of these two variables is reflected in the fact that for most schools neither the estimated proportion of students meeting the standard nor the comparison with the statewide average changed markedly. The exceptions to this lack of impact were again schools 20 and 30. In particular, when ethnicity and gender were considered, neither school had significantly different proportions of students meeting the standard than the statewide average. Given that neither of these variables were significantly related to students meeting the growth standard for reading, it is also possible that their inclusion in the model resulted in nonsignificant results for these schools due in part to a statistical anomaly caused by the use of degrees of freedom without a corresponding decline in unexplained variation in the outcome.

10.2 Mathematics

Table 6 includes the hypothesis testing results for the three models with the outcome variable being whether or not the examinee met the expected growth standard for the mathematics exam. For all three models, school was significantly related to the outcome, meaning that there were significant differences in proportions across schools.

Table 6 Significance test results for math exam, independent variable, and school level

Variable	Model 1	Model 2	Model 3
School	Chi-square=540.594 *df*=40 $p<0.001$	Chi-square=512.076 *df*=40 $p<0.001$	Chi-square=480.683 *df*=40 $p<0.001$
Grade		Chi-square=32.897 *df*=1 $p<0.001$ $b=-0.032$	Chi-square=33.289 *df*=1 $p<0.001$ $b=-0.032$
Gender			Chi-square=2.189 *df*=1 $p=0.139$ $b=-0.035$
Caucasian			Chi-square=6.510 *df*=1 $p=0.011$ $b=0.107$

An examination of Tables 7 and 8 reveals that schools 3, 6, 9, 11, 23, 26, 28, 29, 31, and 38 all had significantly higher proportions of examinees meeting the math standard as compared to the state as a whole. Schools 8, 12, 13, 15, 16, 17, 20, 21, 30, 33, and 37 had significantly lower proportions of students meeting the standard, while the rest had rates comparable to the state average. Results for model 2 revealed that grade was inversely related to the likelihood of meeting the standard, as was true for reading. While the adjusted school wide proportions were not generally different for most schools between models 1 and 2, the adjusted results were markedly different for a few of the schools. As an example, school 25 was not found to have significantly different rates than the state as a whole when only school was considered, but when the results were adjusted for grade, it was found to have significantly lower rates, with the actual school estimate going from 0.50 to 0.45. Similarly, school 23 was initially found to have significantly higher rates of students meeting the standard, but when grade was included, its rate declined from 0.62 to 0.60, which was not significantly higher than the statewide rate of 0.57. Further investigation revealed that 76 % of students in school 23 were in grade 3 or lower. Thus, when the model controls for these relatively higher performing grade levels, students in this school were not found to perform significantly better than those across the state. School 25 was comprised entirely of students in grades 1 or 2, so that once again, when these higher performing grades were controlled for, the relative performance of the school as a whole declined vis-à-vis that of the state as a whole.

Finally, as with the reading test, model 3 included school, grade, gender, and ethnicity. Results in Table 6 show that school, grade, and ethnicity were all significantly related to the likelihood that an individual would meet the mathematics growth standard. In terms of the school results, number 23 continued to exhibit a nonsignificant proportion meeting the standard when compared with the state, and number 25 had a significantly lower such rate, as was true for model 2. On the other hand, school 27 went from having a nonsignificant result to having a significantly higher rate of students meeting the mathematics standard than that of the state as a whole, when gender and ethnicity, as well as grade, were accounted for in the model. Indeed, the difference in the estimated rate meeting the standard for the school as compared to the state increased from 0.08 when only school was considered to 0.13 when school, grade, ethnicity, and gender were all included in the model. In addition, the hypothesis test result for school 33 was not significant for model 3, though it should be noted that the p value was still very close to the α of 0.05. Furthermore, the actual proportion meeting the standard for this school did not change, suggesting that the nonsignificant test result was a function of the change in standard error value in the more complex model. Examination of student demographics showed that 100 % of students in school 27 were non-Caucasian, as were 92.3 % of those attending school 33. The statewide value was 79.2 %. Thus, when this disparity in ethnic distribution was controlled, the estimated proportion meeting the standard for both schools improved when compared with that of the entire state, particularly for school 27.

10.3 Language

For the language exam, hypothesis test results in Tables 8 and 9 show that schools were significantly different in terms of the proportion of students meeting the growth

Table 7 Proportion of examinees meeting the growth standard for elementary math test

School	School only	School and grade	School, grade, sex, race
State average	0.57	0.57	0.58
1 (247)	0.57	0.57	0.59
2 (203)	0.54	0.58	0.58
3 (576)	0.83	0.83	0.84
4 (535)	0.57	0.56	0.58
5 (176)	0.50	0.51	0.53
6 (416)	0.65	0.63	0.65
7 (215)	0.59	0.60	0.62
8 (258)	0.45	0.46	0.48
9 (163)	0.72	0.72	0.73
10 (88)	0.64	0.63	0.64
11 (380)	0.67	0.67	0.69
12 (363)	0.51	0.50	0.53
13 (114)	0.33	0.33	0.35
14 (201)	0.54	0.53	0.53
15 (486)	0.51	0.51	0.53
16 (80)	0.44	0.45	0.44
17 (205)	0.28	0.32	0.33
18 (223)	0.55	0.57	0.56
19 (60)	0.57	0.58	0.60
20 (530)	0.47	0.46	0.48
21 (412)	0.41	0.40	0.43
22 (691)	0.58	0.58	0.58
23 (359)	0.62	0.60	0.61
24 (242)	0.53	0.55	0.57
25 (84)	0.50	0.45	0.47
26 (423)	0.64	0.63	0.64
27 (40)	0.65	0.69	0.71
28 (481)	0.68	0.68	0.67
29 (178)	0.70	0.73	0.75
30 (218)	0.39	0.42	0.44
31 (409)	0.67	0.67	0.69
32 (88)	0.65	0.66	0.65
33 (156)	0.48	0.49	0.50
34 (116)	0.62	0.63	0.62
35 (175)	0.62	0.62	0.64
36 (104)	0.62	0.63	0.61
37 (881)	0.48	0.49	0.51
38 (116)	0.66	0.66	0.68
39 (124)	0.63	0.63	0.63
40 (449)	0.60	0.60	0.62
41 (217)	0.57	0.56	0.57

Table 8 Deviance contrast results for elementary school math test: (difference, p value)

School	School only	School and grade	School, grade, sex, race
1 (247)	0.00 (p=0.993)	0.00 (p=0.927)	0.01 (p=0.849)
2 (203)	−0.02 (p=0.476)	0.01 (p=0.767)	0.00 (p=0.890)
3 (576)	0.26 (p=0.000)	0.26 (p=0.000)	0.26 (p=0.000)
4 (535)	0.01 (p=0.802)	−0.01 (p=0.737)	0.00 (p=0.969)
5 (176)	−0.07 (p=0.074)	−0.06 (p=0.133)	−0.05 (p=0.213)
6 (416)	0.08 (p=0.001)	0.06 (p=0.015)	0.07 (p=0.004)
7 (215)	0.02 (p=0.557)	0.03 (p=0.408)	0.04 (p=0.277)
8 (258)	−0.12 (p=0.000)	−0.11 (p=0.001)	−0.10 (p=0.001)
9 (163)	0.15 (p=0.000)	0.15 (p=0.000)	0.15 (p=0.000)
10 (88)	0.07 (p=0.165)	0.06 (p=0.225)	0.05 (p=0.278)
11 (380)	0.10 (p=0.000)	0.10 (p=0.000)	0.10 (p=0.000)
12 (363)	−0.06 (p=0.023)	−0.06 (p=0.014)	−0.06 (p=0.038)
13 (114)	−0.23 (p=0.000)	−0.24 (p=0.000)	−0.23 (p=0.000)
14 (201)	−0.02 (p=0.486)	−0.04 (p=0.209)	−0.05 (p=0.169)
15 (486)	−0.06 (p=0.011)	−0.06 (p=0.008)	−0.05 (p=0.029)
16 (80)	−0.13 (p=0.018)	−0.12 (p=0.028)	−0.15 (p=0.009)
17 (205)	−0.28 (p=0.000)	−0.25 (p=0.000)	−0.25 (p=0.000)
18 (223)	−0.01 (p=0.650)	0.00 (p=0.906)	−0.02 (p=0.536)
19 (60)	0.00 (p=0.998)	0.01 (p=0.881)	0.02 (p=0.784)
20 (530)	−0.10 (p=0.000)	−0.11 (p=0.000)	−0.10 (p=0.000)
21 (412)	−0.15 (p=0.000)	−0.16 (p=0.000)	−0.16 (p=0.000)
22 (691)	0.02 (p=0.384)	0.01 (p=0.569)	0.00 (p=0.867)
23 (359)	0.05 (p=0.033)	0.03 (p=0.265)	0.03 (p=0.318)
24 (242)	−0.03 (p=0.293)	−0.02 (p=0.530)	−0.02 (p=0.629)
25 (84)	−0.07 (p=0.214)	−0.12 (p=0.031)	−0.11 (p=0.046)
26 (423)	0.07 (p=0.002)	0.06 (p=0.006)	0.06 (p=0.010)
27 (40)	0.08 (p=0.258)	0.12 (p=0.087)	0.13 (p=0.031)
28 (481)	0.12 (p=0.000)	0.11 (p=0.000)	0.09 (p=0.000)
29 (178)	0.14 (p=0.000)	0.16 (p=0.000)	0.16 (p=0.000)
30 (218)	−0.18 (p=0.000)	−0.15 (p=0.000)	−0.14 (p=0.000)
31 (409)	0.11 (p=0.000)	0.10 (p=0.000)	0.11 (p=0.000)
32 (88)	0.08 (p=0.104)	0.09 (p=0.074)	0.07 (p=0.189)
33 (156)	−0.09 (p=0.030)	−0.08 (p=0.035)	−0.08 (p=0.051)
34 (116)	0.05 (p=0.222)	0.06 (p=0.201)	0.04 (p=0.396)
35 (175)	0.05 (p=0.163)	0.05 (p=0.169)	0.06 (p=0.104)
36 (104)	0.06 (p=0.210)	0.06 (p=0.204)	0.03 (p=0.585)
37 (881)	−0.08 (p=0.000)	−0.08 (p=0.000)	−0.07 (p=0.000)
38 (116)	0.09 (p=0.041)	0.09 (p=0.033)	0.10 (p=0.018)
39 (124)	0.06 (p=0.143)	0.06 (p=0.181)	0.05 (p=0.257)
40 (449)	0.03 (p=0.135)	0.03 (p=0.167)	0.04 (p=0.077)
41 (217)	0.00 (p=0.883)	−0.01 (p=0.842)	−0.01 (p=0.769)

Table 9 Significance test results for math exam, independent variable and school level

Variable	Model 1	Model 2	Model 3
School	Chi-square=271.972 df=38 p<0.001	Chi-square=272.794 df=38 p<0.001	Chi-square=264.753 df=38 p<0.001
Grade		Chi-square=35.828 df=1 p<0.001 b=0.048	Chi-square=35.817 df=1 p<0.001 b=0.048
Gender			Chi-square=4.782 df=1 p=0.029 b=−0.059
Caucasian			Chi-square=0.075 df=1 p=0.784 b=0.013

standard. An examination of Tables 10 and 11 reveals that schools 3, 24, 28, 31, 35, and 38 all had significantly higher proportions than did the state as a whole, while schools 13, 20, 21, 30, and 37 had significantly lower proportions. Also, note that schools 17 and 25 did not have any students participating in both the fall and spring language examination. Grade (model 2) was also significantly related to whether students met their target growth rates, displaying a positive relationship, unlike for the other two exams. In other words, for the language exam, students in higher grades had a greater likelihood of meeting the expected growth standard. Under model 2, school 11 had a significantly higher proportion meeting the standard than the statewide value, while school 21 went from having a significantly lower rate to not being significantly different from the statewide average. It was found that both of these schools had a preponderance of students in the lower grade levels, which when controlled for made their relative performance better. Results for model 3 showed that in addition to school and grade, gender, but not ethnicity, was significantly related to the likelihood of an examinee meeting the growth standard. In terms of relative school comparisons, there were no substantive changes beyond those described for model 2.

11 Conclusions

The goal of this research was to demonstrate the utility of the probit model for providing estimates of school level proportions of students meeting an achievement test performance standard and adjusting these estimates for salient variables known to impact achievement. This model, which assumes that the dichotomous outcome of interest is based on an underlying continuum, seems very well suited to the analysis of standard data drawn from educational achievement testing programs such as those used by most states. The probit model provides estimates of the proportion of examinees meeting the standard for each school, and by adding covariates of interest to the model,

Table 10 Proportion of examinees meeting the growth standard for elementary language test

School	School only	School and grade	School, grade, sex, race
State average	0.56	0.56	0.56
1 (247)	0.58	0.57	0.57
2 (203)	0.61	0.57	0.57
3 (576)	0.74	0.74	0.74
4 (535)	0.56	0.57	0.57
5 (176)	0.49	0.48	0.49
6 (416)	0.51	0.53	0.53
7 (215)	0.54	0.54	0.54
8 (258)	0.51	0.50	0.50
9 (163)	0.57	0.59	0.59
10 (88)	0.61	0.64	0.64
11 (380)	0.64	0.65	0.65
12 (363)	0.58	0.57	0.57
13 (114)	0.27	0.28	0.28
14 (201)	0.55	0.57	0.57
15 (486)	0.52	0.52	0.52
16 (80)	0.57	0.57	0.57
17 (0)			
18 (223)	0.55	0.54	0.54
19 (60)	0.53	0.52	0.53
20 (530)	0.45	0.47	0.47
21 (412)	0.46	0.47	0.47
22 (691)	0.52	0.52	0.52
23 (359)	0.47	0.50	0.50
24 (242)	0.69	0.69	0.69
25 (0)			
26 (423)	0.58	0.59	0.59
27 (40)	0.41	0.36	0.36
28 (481)	0.65	0.65	0.65
29 (178)	0.57	0.54	0.54
30 (218)	0.40	0.37	0.37
31 (409)	0.66	0.66	0.67
32 (88)	0.57	0.57	0.57
33 (156)	0.51	0.52	0.52
34 (116)	0.55	0.56	0.56
35 (175)	0.72	0.71	0.72
36 (104)	0.68	0.68	0.68
37 (881)	0.48	0.46	0.47
38 (116)	0.74	0.75	0.75
39 (124)	0.58	0.58	0.58
40 (449)	0.55	0.55	0.55
41 (217)	0.57	0.59	0.59

Table 11 Deviance contrast results for elementary school language test: (difference, *p* value)

School	School only	School and grade	School, grade, sex, race
1 (247)	0.02 (p=1.000)	0.01 (p=1.000)	0.01 (p=1.000)
2 (203)	0.05 (p=0.998)	0.01 (p=1.000)	0.01 (p=1.000)
3 (576)	0.18 (p=0.000)	0.18 (p=0.000)	0.18 (p=0.000)
4 (535)	0.00 (p=1.000)	0.01 (p=1.000)	0.02 (p=1.000)
5 (176)	−0.07 (p=0.924)	−0.07 (p=0.845)	−0.07 (p=0.894)
6 (416)	−0.05 (p=0.989)	−0.02 (p=1.000)	−0.02 (p=1.000)
7 (215)	−0.02 (p=1.000)	−0.02 (p=1.000)	−0.02 (p=1.000)
8 (258)	−0.05 (p=0.989)	−0.06 (p=0.957)	−0.05 (p=0.965)
9 (163)	0.01 (p=1.000)	0.03 (p=1.000)	0.03 (p=1.000)
10 (88)	0.06 (p=1.000)	0.08 (p=0.992)	0.08 (p=0.994)
11 (380)	0.09 (p=0.072)	0.09 (p=0.049)	0.09 (p=0.048)
12 (363)	0.02 (p=1.000)	0.01 (p=1.000)	0.01 (p=1.000)
13 (114)	−0.29 (p=0.000)	−0.28 (p=0.000)	−0.28 (p=0.000)
14 (201)	−0.01 (p=1.000)	0.01 (p=1.000)	0.01 (p=1.000)
15 (486)	−0.03 (p=1.000)	−0.03 (p=1.000)	−0.03 (p=1.000)
16 (80)	0.02 (p=1.000)	0.02 (p=1.000)	0.01 (p=1.000)
17 (0)	NA	NA	NA
18 (223)	−0.01 (p=1.000)	−0.02 (p=1.000)	−0.02 (p=1.000)
19 (60)	−0.03 (p=1.000)	−0.03 (p=1.000)	−0.03 (p=1.000)
20 (530)	−0.11 (p=0.001)	−0.09 (p=0.018)	−0.09 (p=0.029)
21 (412)	−0.10 (p=0.013)	−0.09 (p=0.073)	−0.09 (p=0.098)
22 (691)	−0.04 (p=0.916)	−0.04 (p=0.985)	−0.04 (p=0.981)
23 (359)	−0.09 (p=0.206)	−0.06 (p=0.942)	−0.06 (p=0.942)
24 (242)	0.14 (p=0.001)	0.13 (p=0.004)	0.13 (p=0.004)
25 (0)	NA	NA	NA
26 (423)	0.03 (p=1.000)	0.03 (p=1.000)	0.03 (p=1.000)
27 (40)	−0.15 (p=0.892)	−0.20 (p=0.290)	−0.20 (p=0.297)
28 (481)	0.09 (p=0.011)	0.09 (p=0.008)	0.09 (p=0.029)
29 (178)	0.01 (p=1.000)	−0.02 (p=1.000)	−0.01 (p=1.000)
30 (218)	−0.15 (p=0.000)	−0.18 (p=0.000)	−0.18 (p=0.000)
31 (409)	0.10 (p=0.010)	0.11 (p=0.006)	0.11 (p=0.005)
32 (88)	0.01 (p=1.000)	0.01 (p=1.000)	0.01 (p=1.000)
33 (156)	−0.05 (p=1.000)	−0.04 (p=1.000)	−0.04 (p=1.000)
34 (116)	−0.01 (p=1.000)	0.00 (p=1.000)	0.00 (p=1.000)
35 (175)	0.16 (p=0.001)	0.16 (p=0.003)	0.16 (p=0.003)
36 (104)	0.12 (p=0.400)	0.12 (p=0.384)	0.12 (p=0.511)
37 (881)	−0.08 (p=0.007)	−0.09 (p=0.000)	−0.09 (p=0.000)
38 (116)	0.19 (p=0.000)	0.19 (p=0.000)	0.19 (p=0.000)
39 (124)	0.02 (p=1.000)	0.02 (p=1.000)	0.02 (p=1.000)
40 (449)	−0.01 (p=1.000)	0.00 (p=1.000)	0.00 (p=1.000)
41 (217)	0.02 (p=1.000)	0.03 (p=1.000)	0.03 (p=1.000)

researchers can also obtain adjusted school level proportions that take into account salient demographic or other features of examinees. Such results may be particularly useful to policy makers and others who are interested in school performance comparisons when such characteristics of the school are taken into account. These comparisons are potentially more useful for educational policy makers and others because they may, in some sense, better reflect the relative performance of the schools than do simple proportions, given that they control for demographic differences that are beyond institutional control and that impact student performance. Furthermore, this approach to estimating aggregated school performance as the proportion of examinees meeting a particular testing standard is easy to carry out using the probit model, which can be fit using standard statistical software such as SPSS. In addition, the results are in the familiar metric of proportion meeting the standard, making them easy for all constituents (e.g., parents, teachers, administrators, policy makers) to understand. In addition, the proportion of examinees meeting the educational standard can be estimated using models that account for relevant demographic factors, as well as models excluding such factors, and the results compared. If the results for a given school differ when these variables are, and are not included in the analysis, this is evidence that the variables do impact student and school performance. In short, when achievement testing results are reported in terms of examinees meeting or not meeting a particular standard, the probit model provides researchers and policy makers with an easy to use tool that produces results in a familiar metric (proportion meeting the standard), while statistically controlling for variables believed to be relevant to student and school performance. Such results can then be compared across schools or within the same school accounting for and not accounting for the control variables. Finally, using the deviation contrast with Sidak's adjustment to control the type I error rate, we were able to compare the proportion meeting the standard in each school with the statewide proportion, thus providing a mechanism for identifying schools that performed above, at, or below average in terms of meeting the standard. It should be noted that an additional advantage of the probit model is that it not only provides adjusted proportions of students meeting the standard of interest, but also adjusted standard errors (precision) for these proportions. Taken together, we view this use of the probit model and accompanying contrasts as a powerful tool for researchers, policy makers, and educators in understanding relative school performance. In particular, the probit model has the potential to provide a greater depth of understanding regarding differences in school academic performance, as well as a more accurate estimate of such differences. Through statistical consideration of variables that are known to influence student academic achievement, but which are out of the control of schools (e.g., student grade, gender, socioeconomic status), the probit model can be used to provide estimates of school performance, in the context of proportion of examinees meeting a standard, that more clearly reflect school effects, above and beyond the impact of these factors beyond school control. The probit model also provides more in depth explanation of school differences in academic performance through estimation of effects of additional variables of interest, and the comparison of differences in school effects for models with and without these variables, as was demonstrated above.

With respect to the specific results of this study, we found that for these charter schools, grade level was a consistently important factor in terms of student attainment of the achievement growth standard. For reading and math, those in higher grades were less likely to meet the standard, while the opposite was true for the language test. Furthermore, when grade was accounted for in the probit model, the estimate proportion meeting the standards and the relative comparison to the statewide performance changed for several schools. In particular, institutions that had a greater ratio of students in the upper grades fared relatively better in reading and math when grade level was included in the model than when it was not. Conversely, schools with a preponderance of examinees in the lower grades saw their relative performance decline when it was controlled for in the model. The opposite pattern was in evidence for the language exam. The student demographic characteristics of gender and ethnicity were not as reliably associated with the proportion of examinees meeting the standard. Gender was significantly related to attainment of the language standard, while ethnicity was related to that for math. Again, controlling for each variable when it was associated with the outcome led to somewhat different corrected results for a small number of schools than when these variables were not accounted for.

11.1 Study limitations

While overall the probit model performed well in terms of providing adjusted school proportions, no method or study is without flaws. For example, the probit model does work under the assumption that there exists an underlying continuum that drives the outcome variable of interest, in this case whether examinees met the standard or not. Furthermore, it is assumed that this continuum follows a normal distribution. For the current study, and indeed much educational research dealing with standardized achievement tests, the assumption of an underlying continuum is probably reasonable. Most standard setting scenarios involve using the results of standardized tests to determine whether an individual has met a particular performance standard, thus meeting the assumption of the dichotomous outcome being based on an underlying continuum. On the other hand, it may not always be the case that this continuum is normally distributed. This may be particularly true in charter schools, which often attract students from either end of the achievement spectrum, creating the potential for skewed data distributions. However, in the current situation with the very large sample sizes involved, the assumption was found to be tenable. Nonetheless, this is an issue to which researchers need to be sensitive when using the probit model, particularly with smaller samples. And, indeed, an assessment of the data revealed that growth from fall to spring for all three tests was in fact normally distributed. However, this may not be the case in all situations, for instance when used for decision making at the individual school level or in small districts. Future study in this area should examine the performance of the probit model in cases where this normality assumption has not been met. In addition, future studies could compare the relative performance of the VAM and logistic models (e.g., Choi and Goldschmidt 2012) with that of the probit to ascertain whether any differences are in evidence with respect to their ability to provide estimates and adjustments to the proportion of examinees meeting the standard. Finally, the current study used MAP data from a group of charter schools. Though the probit model described here is easily applied to, and very appropriate for any educational context in which the outcome variable is a dichotomy, such as whether students have met an academic standard or not, we do recognize that charter schools are different in some important respects from other public schools. In particular, they typically have somewhat greater freedom in terms of the qualifications of those hired as teachers, as well as in setting the length of the academic calendar. At the same time, they must meet the academic requirements that are expected of all public schools in the state. In addition, the MAP is only one method of assessing student performance and is not used by all schools or school systems. Future research in this area should expand the use of the probit model for assessing school performance to contexts beyond charter schools, and to situations in which a dichotomous outcome variable measures student performance in a different way, such as whether a particular score level was achieved on a single test administration, as opposed to the growth standard used here.

References

Agresti, A. (2002). *Categorical data analysis*. New York: Wiley.

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., and Zhang, J. (2013). *The Condition of Education 2013 (NCES 2013-037)*: U.S. department of education, national center for education statistics. Washington, DC. http://nces.ed.gov/pubsearch.

Azen, R., & Walker, C. M. (2011). *Categorical data analysis for the behavioral and social sciences*. New York: Routledge.

Baker, M., & Johnston, P. (2010). The impact of socioeconomic status on high stakes testing reexamined. *Journal of Instructional Psychology, 37*(3), 193–199.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 21*, 37–66.

Barton, P. E. (2008). The right way to measure growth. *Educational Leadership, 65*, 70–73.

Braun, H. (2005). *Using student progress to evaluate teachers: a primer to value-added models*. Princeton: ETS.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy, 4*(4), 385–414.

Capraro, R. M., Young, J. R., Lewis, C. W., Yetkiner, Z. E., & Woods, M. N. (2009). An examination of mathematics achievement and growth in a midwestern urban school district: implications for teachers and administrators. *Journal of Urban Mathematics Education, 2*(2), 46–65.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*, 1045–1057.

Choi, K., & Goldschmidt, P. (2012). A multilevel latent growth curve approach to predicting student proficiency. *Asia Pacific Education Review, 13*(2), 199–208.

IBM Corp. (2010). *IBM SPSS Statistics for Windows, version 19.0*. Armonk: IBM Corp.

Darling-Hammond, L., Amerin-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.

De Lisle, J., Smith, P., & Jules, V. (2010). Evaluating the geography of gendered achievement using large-scale assessment data from the primary school system of the Republic of Trinidad and Tobago. *International Journal of Educational Development, 30*(4), 405–417.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are "failing" schools really failing? Using seasonal comparison to evaluate school effectiveness. *Sociology of Education, 81*(3), 242–270.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks: Sage.

Franco, M. S., & Seidel, K. (2012). Evidence for the need to more closely examine school effects in value-added modeling and related accountability policies. *Education and Urban Society*. doi:10.1177/0013124511432306.

Goldschmidt, P., Roschewski, P., Choi, L., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). Policymakers' guide to growth models for school accountability: How do accountability models differ? Washington DC: The Council of Chief State School Officers. http://www.ccsso.org/Resources/Publications/Policymakers%E2%80%99_Guide_to_Growth_Models_for_School_Accountability_How_Do_Accountability_Models_Differ.html. Accessed 14 Jan 2014.

Goldschmidt, P., Choi, K., Martinez, F., & Novak, J. (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment. *School Effectiveness and School Improvement: An International Journal of Research, Policy, and Practice, 21*(2), 337–357.

Gorard, S. (2011). Now you see it, now you don't: school effectiveness as conjuring? *Research in Education, 86*, 39–45.

Lee, J. (2010). Tripartite growth trajectories of reading and math achievement: tracking national academic progress at primary, middle, and high school levels. *American Educational Research Journal, 47*(4), 800–832.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*, 4–16.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, F. (2006). *The sensitivity of value-added teacher effect estimates to different mathematics achievement measures*. Santa Monica: RAND.

Martineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement data for value-added accountability. *Journal of Educational and Behavioral Statistics, 31*, 35–62.

McCaffrey, D.F. (2013). Do value-added methods level the playing field for teachers? *Carnegie Knowledge Network.* http://carnegieknowledgenetwork.org/briefs/value-added/level-playing-field/.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Northwest Evaluation Association. (2003). *Technical manual for use with measures of academic progress and achievement level tests*. Portland: Northwest Evaluation Association.

Olson, L. (2004). Value added models gain in popularity. *Education Week, 24*(12), 14–15.

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112*(4), 1137–1162.

Scherrer, J. (2012). What's the value of VAM (value-added modeling)? *Phi Delta Kappan, 93*(8), 58–60.

Schmidt, W. H., Houang, R. T., & McKnight, C. C. (2005). Value-added research: right idea but wrong solution? In R. Lissitz (Ed.), *Value added models in education: theory and practice* (pp. 272–297). Maple Grove: JAM.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of American Statistical Association, 67*(62), 626–633.

Tong, H., & Lim, K. S. (1980). Threshold autoregression, limit cycles, and cyclical data. *Journal of the Royal Statistical Society, Series B, 42*, 245–292.

Wainer, H. (2000). *Computer adaptive testing: a primer.* Mahwah: Lawrence Erlbaum.

Weiss, M. J., & May, H. (2012). A policy analysis of the federal growth model pilot program's measures of school performance: the Florida case. *Association for Education Finance and Policy, 7*(1), 44–73. doi:10.1162/EDFP_a_00053.

Wiggan, G. (2007). Race, school achievement, and educational inequality: toward a student-based inquiry perspective. *Review of Educational Research, 77*(3), 310–333.