METHODOLOGY

OVERVIEW

Measure of Change Over Time

Inclusion rates can vary among states and across time owing to[4]

- differing proportions of students with different types and severities of disability;
- differing accommodations offered by the states for their own state assessment tests;
- measures taken by NCES to increase the number of students with disabilities who are included; and
- other factors not associated with characteristics of the states' SD population or policies for accommodations on their own state assessment tests.

The motivation behind this report is that state-level inclusion rates are *expected* to vary according to differing proportions of students with different types and severities of disabilities and the offering of accommodations on the state assessment that are not allowed on NAEP. Variations that result from other factors that we cannot measure, such as actions taken by NCES, are not standard and are meant to be captured by our change measure. This breakdown lends itself to an analogy to studies that attempt to measure discrimination. In the discrimination case, wages are *expected* to vary according to certain demographic characteristics, such as education and experience. However, wages can also vary because of factors we cannot measure, such as discrimination. Studies of discrimination have commonly used the Oaxaca-Blinder technique to decompose differences in wages into a portion that is expected and a portion that is not. The similarities to the discrimination application motivated us to borrow from Oaxaca-Blinder decomposition for the development of our methodology for measuring change.

The Oaxaca-Blinder decomposition technique partitions the difference of the means between two groups into a portion explained by differences in control variables and a portion that is explained by differences in how those characteristics are treated/rewarded. In the discrimination application, it is the difference in the mean wages between women and men, for example. In our application, it is the difference between a state's inclusion rate in the initial period, 2005, and in the second period, 2007. Because our focus is on a state-by-state analysis and not on a national analysis, we need to apply the technique 51 times: one for each state and the District of Columbia.

Both the Oaxaca-Blinder decomposition technique and our partitioning technique measure the portion of mean group differences attributed to differences in underlying characteristics by fixing the individual-level relationship between observed characteristics and outcome. In both techniques, this relationship is held constant across groups being compared. In the application of our partitioning technique, as described further below, this fixed relationship acts as the yardstick for comparison.

We use Fairlie's (2003) framework for Oaxaca-Blinder decomposition in the non-linear case to explain the differences and similarities with our partitioning technique. Fairlie provides the

4 A discussion of how these factors affect inclusion rates is provided later in this report.

following equation for the Oaxaca-Blinder decomposition of different outcomes between blacks and whites in a non-linear case:[5]

$$\overline{Y}^W - \overline{Y}^B = \left[\sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^W)}{N^W} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^W)}{N^B}\right] + \left[\sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^W)}{N^B} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^B)}{N^B}\right],$$

where

$\overline{Y}^W - \overline{Y}^B$ is the difference in overall mean outcome between whites and blacks, respectively;

$F(\bullet)$ is a non-linear function;

X_i^W and X_i^B are vectors of control variables for whites and blacks, respectively;

N^W and N^B are the number of observations for whites and blacks, respectively; and

$\hat{\beta}^W$ and $\hat{\beta}^B$ are the vectors of coefficients from the non-linear regressions estimated for whites and blacks, respectively.

The term in the first set of brackets is the portion of the difference in overall means that is due to different distributions of the control variables while the second set of brackets contains the portion that is due to differences in overall group average outcome.

Our partitioning technique diverges from the Oaxaca-Blinder decomposition by simply subtracting out the portion in the first bracket from the difference in overall means and using the remainder as our *change measure*.

$$Change = \left[\overline{Y}^2 - \overline{Y}^1\right] - \left[\sum_{i=1}^{N^2} \frac{\Lambda(X_i^2 \hat{\beta}^*)}{N^2} - \sum_{i=1}^{N^1} \frac{\Lambda(X_i^1 \hat{\beta}^*)}{N^1}\right]$$

Where:

$\overline{Y}^2 - \overline{Y}^1$ is the difference in overall mean outcome between periods 2 and 1, respectively;

$\Lambda(z) = \dfrac{e^z}{1+e^z}$ is the logistic cumulative distribution function;

X_i^2 and X_i^1 are vectors of control variables for period 2 and period 1, respectively;

N^2 and N^1 are number of observations for period 2 and period 1, respectively; and

$\hat{\beta}^*$ is a vector of regression coefficients.

In both the Oaxaca-Blinder decomposition and our partitioning technique, the relationship between outcome and controls, $\hat{\beta}^W$ and $\hat{\beta}^B$ in the Oaxaca-Blinder equation and $\hat{\beta}^*$ in our partitioning equation, is estimated by using regression analysis at the individual level. Within the Oaxaca-Blinder framework and our partitioning technique, how these coefficients are derived can vary. In particular, the population on which this relationship is estimated can vary and will, hence, provide slightly different measures.

In this vein, we developed two approaches to applying our partitioning technique for measuring change in state-level inclusion rates over time: the nation-based approach and the

5 Fairlie (2003), p 2.

state-specific approach.[6] The two approaches differ in how the relationship between inclusion on NAEP and student characteristics, $\hat{\beta}^*$, is estimated:

- In the nation-based approach, one regression on initial-period national data is used to fix the relationship between inclusion on NAEP and SD characteristics. Change in each state is measured using that same estimated relationship.
- In the state-specific approach, a regression model is estimated separately for each state to fix the relationship between inclusion on NAEP and student characteristics. The regression is estimated for each state using that state's SD sample in the initial year.

Once the relationship is fixed in the form of the estimated coefficients, it is applied to the states' data (to the initial and second period data under the nation-based approach; to the second period data alone in the state-specific approach) to provide individual-level predicted probabilities of inclusion for each student. The predicted probabilities for each student are based on his or her control characteristics. Within each state in each time period, the student-level predicted probabilities are aggregated to the state level to provide a state-level predicted inclusion rate. As an aggregation of student-level predicted probabilities, the predicted inclusion rate for each state in each period is based on the state's distribution of student characteristics. The predicted inclusion rates for each state in each time period is then compared with the actual inclusion rate, and the differences are then used to construct the change measure as in the discrimination examples above. Exact details and formulas for estimation, aggregation, and measure determination are provided below separately for each approach.

Measure of Starting Point

In addition to providing the measures of change in inclusion rates over time described above, we provide a context for this change by comparing states' inclusion rates on NAEP in the initial period. We refer to this measure as the *starting point* for each state. Even when we hold constant different types and severities of disabilities and different accommodations offered by the states for their own state assessments, not all states start with the same inclusion rate for NAEP. In some states, SDs are initially included at higher rates than in others; therefore, we would expect less change in including students in these states.

Each approach to measuring change is discussed below, and the discussion includes explanations of how the measure of *starting point* is calculated. The use of the starting point measure vis-à-vis the change measure is also discussed in detail below.

A separate starting point measure was developed for each of the nation-based and state-specific approaches to measuring change. Those different starting point measures were designed for the approaches under which they were developed, but both can be used with either approach. We present them without preference because each has its benefits and drawbacks.

- In the nation-based approach, the same estimated regression model used to fix the relationship between inclusion on NAEP and SD student characteristics is also used to calculate differences between states in the initial period. In other words, we use previously calculated results to construct the starting point measure.
- In the state-specific approach, the regressions estimated to calculate the change measure are state specific and cannot be used to compare states or provide a starting point measure. For this reason, a separate regression model is estimated on the entire NAEP SD

6 The terms nation-based and state-specific describe the different approaches and, in particular, how the relationship between inclusion and SD characteristics is fixed. All analysis of change is done on a state-by-state basis.

sample in the initial period to generate a starting point measure. The model is the same as that used under the nation-based approach but includes state fixed effects.

Building on the Oaxaca-Blinder methodology used to measure change, we similarly fix the relationship between control and outcome variables to provide individual-level predicted probabilities. The individual-level predicted probabilities serve as a basis of comparison. Details and formulas for estimation, aggregation, and measure determination are provided below separately for each approach.

NATION-BASED APPROACH

In the nation-based approach, one regression on national data is used to fix the relationship between inclusion on NAEP and SD student characteristics. Here, the entire NAEP sample for the initial year is used to estimate the relationship between student characteristics and the probability of inclusion with no differences between states explicitly modeled. The estimated coefficients are applied to each year of data to provide a predicted probability of inclusion for each SD. The *average* predicted probability of inclusion for all students with a disability in a given state in a given year is then the benchmark for that state for that year, or, in other words, that state's predicted inclusion rate for that year. Whereas the predicted probability for a student with a given set of characteristics is fixed by the model and does not change across states or time, the predicted inclusion rate for each state is different and changes across time because of differences in the populations of SDs.

In practice, the nation-based approach is based on the following regression model:

$$Included_i = \Lambda\left(\alpha + \sum_j \sum_l \sum_k \left(\beta_{j,l,k} \cdot DisabilityType_i^j \cdot SeverityType_i^k \cdot GradeLevel_i^l\right) + \varphi \cdot NonNaepAcc_i\right),$$

where

$Included_i$ = 1 if student i was included on the NAEP assessment; 0 otherwise;

$\Lambda(z) = \dfrac{e^z}{1+e^z}$ is the logistic cumulative distribution function;

$DisabilityType_i^j$ = 1 if student i has disability type j; 0 otherwise;

$SeverityType_i^k$ = 1 if student i has disability severity level k; 0 otherwise;

$GradeLevel_i^l$ = 1 if student i is receiving instruction at grade level l; 0 otherwise;

$NonNaepAcc_i$ = 1 if student i receives an accommodation on state assessments not allowed on NAEP; 0 otherwise;

$\alpha, \beta_{j,l,k}, \varphi$ are coefficients to be estimated; and

i indexes students, j indexes disability types, k indexes severity levels, and l indexes grade levels of instruction.

This logistic regression is estimated using initial period student-level data and respective sampling weights.[7] The interpretation of this regression is that it provides the *average rate of inclusion* for students with a given set of characteristics in the initial period across the nation. These averages then become our yardstick for measuring state-level changes in inclusion.

7 All estimations and aggregations to state-level statistics use the individual NAEP weights assigned to the data. For details on the use of weights see appendix A.

Included in the model are indicators for the student's type of disability, severity level of disability, and grade level.[8] Each of these measures is crossed with the others so that there is a unique indicator variable for each disability, severity level, and grade level combination. Different disabilities are more or less easy to accommodate on NAEP assessments, and some disabilities hinder learning more than others. Students with disabilities that are classified as severe are expected to be included less often than students whose disabilities are classified as moderate or mild. Grade level of instruction is also an indicator of how severe the disability is. The measure of grade level of instruction is measured on a more objective scale than severity level and additionally is subject specific (mathematics or reading). Also part of the model is an indicator for whether the student received an accommodation on his or her state assessment that was not allowed on NAEP. Students who receive an accommodation on their state assessment that is not allowed on NAEP are expected to be included less often, other things being equal, because the respondent to NAEP's SD Background Questionnaire may judge NAEP's accommodations to be inadequate for the student in question. (See the Data section for further information on the questionnaire.)

Under the nation-based approach, change over time is measured by the change in the difference between the actual inclusion rate and the predicted inclusion rate. State-level actual and predicted inclusion rates are calculated as follows:[9]

$$StateLevelPredicted_s^y = \frac{1}{N_s^y}\sum_{i=1}^{N_s^y} \hat{Included}_i^0 \cdot Weight_i \,,$$

$$StateLevelActual_s^y = \frac{1}{N_s^y}\sum_{i=1}^{N_s^y} Included_i^y \cdot Weight_i \,,$$

$$DistAbovePredicted_s^y = StateLevelActual_s^y - StateLevelPredicted_s^y \,,$$

where

$\hat{Included}_i^0$ is the predicted probability of inclusion for student i, based on the initial period (time=0) model;

$Included_i^y$ = 1 if student i was included on the NAEP assessment in time period y; 0 otherwise;

$Weight_i$ = sampling weight for student *i*;

N_s^y is the sum of weights of all students with disabilities in state s at time period y; and

i indexes students, s indexes states, and y indexes time period (initial=0, second=1).

Change over time for a state is measured by the change in the *distance above the predicted* measure:

$$Change_s = DistAbovePredicted_s^1 - DistAbovePredicted_s^0$$

As an example, if a state's initial-period actual inclusion rate is 3 percentage points above its initial-period predicted inclusion rate (*distance above predicted* in initial period is 3 percentage points), and its second period actual inclusion rate is 5 percentage points above its second-period predicted inclusion rate (*distance above predicted* in initial period is 3 percentage

8 See table 7 for a list of disabilities, severity levels, and grade levels used in the analysis.

9 State-level predicted inclusion rates and distance above the predicted inclusion rate measures are essentially based on average inclusion rates across the country. In preparation of this study, we explored presenting recentered distance above the benchmark measures. A discussion of the rationale along with the recentered results are presented in appendix C.

points), this is an improvement of 2 percentage points. It would also be an improvement of 2 percentage points if a state's initial-period actual inclusion rate is 4 percentage points *below* its initial-period state-level benchmark inclusion rate and its second-period actual inclusion rate is 2 percentage points *below* its second-period benchmark inclusion rate. Improvement is, therefore, movement upward relative to the state-level benchmark inclusion rate and can be an increase in the distance above the benchmark, a decrease in the distance below the benchmark, or movement from below the benchmark to above the benchmark.

Starting Point

To provide the context for the change measure, we compare states in the initial period, the starting point, under each approach. The rationale for providing a starting point measure is that states that initially have a high relative inclusion rate have less room for improvement than states that have a relatively lower inclusion rate to begin with. Hence, a measure of how states compare in relative inclusiveness is a useful context for helping understand the change measure. If a state has a high relative inclusion rate in the starting period, we would not expect a positive change measure. The starting point measure is useful only for comparing states in the period under consideration over which change is measured.

In the nation-based approach, the starting point measure for a state is simply that state's initial-period *distance above predicted* inclusion rate. For example, if State X has an initial *distance above predicted* of −1.1 and State Y has an initial-period *distance above predicted* of −5.5, we conclude that State X has a higher starting point measure than State Y.

$$StartingPoint_s = DistAbovePredicted_s^0$$

STATE-SPECIFIC APPROACH

In the state-specific approach for measuring change, the regression that estimates the relationship between a student's characteristics and the probability of inclusion is calculated separately for each state in the initial period, providing a (potentially) unique yardstick for measurement for each state. Because the regression model is estimated for each state using that state's data in the initial period, it will produce student-level predicted probabilities for that state *in the initial period* that will exactly return the state's actual inclusion rate of that state in the initial period when aggregated to the state level. Hence, to measure change, we need only apply the estimated student-level predicted probabilities to the second year of data.[10]

The intuition behind the state-specific approach is that change in each state is measured relative to itself because a separate yardstick is set up for each state on the basis of that state's initial period data. The predicted probabilities of inclusion for different types of SDs estimated using the initial period data in State X are used as expectations for inclusion of different types of SDs in State X in the second period. The relationship between inclusion and student characteristics that is used as a yardstick is not set by national averages, as it is in the nation-based approach, but is set separately for each state by its own state averages. Unlike the model in the nation-based approach, this model does not include a control for students who receive an accommodation on their state assessment that is not allowed on NAEP. This omission is discussed further below.

10 In other words, were we to apply the student-type benchmarks to the data on which they were estimated, we would end up with a state-level benchmark for the initial period that exactly equaled the state-level actual rate of inclusion for the initial period.

This model used in this approach is as follows:[11]

$$Included_{i,s} = \Lambda\begin{pmatrix} \alpha_s + \sum_j \sum_k \beta_{j,k,s} \cdot DisabilityType_{i,s}^j \cdot SeverityType_{i,s}^k \\ + \sum_l \gamma_{l,s} \cdot GradeLevel_{i,s}^l \end{pmatrix} \text{, for each state } s$$

where

$Included_{i,s}$ = 1 if student i in state s was included on the NAEP assessment; 0 otherwise;

$\Lambda(z) = \frac{e^z}{1+e^z}$ is the logistic cumulative distribution function;

$DisabilityType_{i,s}^j$ = 1 if student i in state s has disability type j; 0 otherwise;

$GradeLevel_{i,s}^l$ = 1 if student i in state s is receiving instruction at grade level l; 0 otherwise;

$SeverityType_{i,s}^k$ = 1 if student i in state s has disability severity level k; 0 otherwise;

α_s, $\beta_{j,k,s}$ and $\gamma_{l,s}$ are coefficients to be estimated; and

i indexes students, j indexes disability types, k indexes severity levels, l indexes grade levels of instruction, and s indexes states.

This logistic regression is estimated separately for each state using that state's initial period student-level data and respective sampling weights. For each state, this estimated model is applied to the state's second year of data to provide a predicted probability of inclusion for each student with a disability in that state in the second period. This predicted probability for each student is based on that student's characteristics. Because the model is estimated separately for each state, each state will (potentially) have a different predicted probability for any given set of student characteristics. The state-level predicted inclusion rate for a state for the second year is the average of the predicted probabilities of inclusion for all students with disabilities in the state in the second year. The measure of change over time is the difference between a state's actual second-period inclusion rate and its state-level benchmark inclusion rate.

In this state-specific approach, we use the same formulas for aggregation and measure construction as in the nation-based approach. However, the formulas can be simplified here because the relationship between inclusion and the control characteristics was estimated using data from each state's initial-year data only. Therefore, in the initial period, the state's predicted inclusion rate will exactly equal the state's actual inclusion rate.

$$StateLevelPredicted_s^0 = \frac{1}{N_s^0} \sum_{i=1}^{N_s^0} \hat{I}ncluded_i^0 \cdot Weight_i$$

$$= \frac{1}{N_s^0} \sum_{i=1}^{N_s^0} Included_i^0 \cdot Weight_i$$

11 An alternative model for the state-specific model was developed but not pursued in this report. Under the alternative model, all data were pooled for estimation of a random coefficients logit model that estimated separate coefficients for each state. Results were very similar to those using the model presented here that estimates a logistic regression separately for each state.

$$= StateLevelActual_s^0$$

Hence, the distance above the predicted inclusion rate for initial period is zero.

$$DistAbovePredicted_s^0 = StateLevelActual_s^0 - StateLevelPredicted_s^0 = 0$$

The change measure then reduces to only the distance above the predicted inclusion rate for the second period.

$$\begin{aligned} Change_s &= DistAbovePredicted_s^1 - DistAbovePredicted_s^0 \\ &= DistAbovePredicted_s^1 - 0 \\ &= DistAbovePredicted_s^1 \end{aligned}$$

As an example, if a state's actual second-period inclusion rate is 2 percentage points above its state-level predicted inclusion rate (i.e., the inclusion rate predicted by the initial-period model), this is considered an improvement of 2 percentage points. If, instead, a state is 3 percentage points *below* its predicted inclusion rate in the second period, this is a 3 percentage point decline in the rate of inclusion.

Starting Point

In the nation-based approach, the relationship between student characteristics and the probability of inclusion, or yardstick, was the same across all states. Hence, it was possible to turn this into a comparison of states in the initial period, 2005. In the state-specific approach, however, the relationship between student characteristics and the probability of inclusion used for measuring change across time is specific to each state and cannot be used to make comparisons among states at any given time. In other words, measurements using different yardsticks cannot be compared. Hence, a second regression, estimated on the sample including all states in the initial period and their respective weights, is used to make comparisons among states:

$$Included_i = \Lambda\left(\begin{array}{l} \alpha + \sum_j \sum_l \sum_k \left(\beta_{j,l,k} \cdot DisabilityType_i^j \cdot SeverityType_i^k \cdot GradeLevel_i^l\right) \\ +\varphi \cdot NonNaepAcc_i + \sum_s \eta_s \cdot State_i^s \end{array}\right),$$

where

$Included_i$ = 1 if student i was included on the NAEP assessment; 0 otherwise;

$\Lambda(z) = \frac{e^z}{1+e^z}$ is the logistic cumulative distribution function;

$DisabilityType_i^j$ = 1 if student i has disability type j; 0 otherwise;

$SeverityType_i^k$ = 1 if student i has disability severity level k; 0 otherwise;

$GradeLevel_i^l$ = 1 if student i is receiving instruction at grade level l; 0 otherwise;

$NonNaepAcc_i$ = 1 if student i receives an accommodation on state assessments not allowed on NAEP; 0 otherwise;

$\alpha, \beta_{j,l,k}, \varphi$, η_s are coefficients to be estimated;

$State_i^s$ = 1 if student i lives in state s; 0 otherwise; and

i indexes students, j indexes types of disability, k indexes severity levels, l indexes grade level of instruction, and s indexes states.

This regression is similar to the nation-based regression but differs because it explicitly estimates differences between states by including state fixed effects or, in other words, indicator variables for each state. If this were a linear model, the state fixed effects could, themselves, be used as the starting point measure because they would be on the same scale as the dependent variable. Because this is a nonlinear model, additional calculations are necessary to translate the fixed effects to the same probability scale as predictions of the dependent variable. Using this second regression, we generate 51 predicted probabilities for each student as if the student were in each state or jurisdiction. The ultimate predicted probability for each student is the average of these 51 predicted probabilities (i.e., the average probability of inclusion across every state). These ultimate predicted probabilities will be the same for each student in the sample, from any state, with the same set of characteristics. This is the common yardstick used to compare states. The state-level predicted inclusion rate is the average of the student-level predicted probabilities in that state. This state-level predicted inclusion rate is interpreted as the *average inclusion rate of all states if all states had the same proportions of students with different types and severities of disabilities as the state in question*. Again, this is performed using initial year data to compare inclusion rates across states in the initial period.

Given the results of the regression equation above, the state-level predicted inclusion rate is as follows:

$$StateLevelPredicted_{\hat{s}} = \frac{\left(\sum_{s=1}^{NumStates} \frac{\sum_{i \in \hat{s}} Prob(inclusion \mid x_i, state = s) \cdot Weight_i}{N_{\hat{s}}} \right)}{NumStates},$$

where

$\hat{s}$ is the reference state;

$N_{\hat{s}}$ is the sum of weights of all students with disabilities in the reference state $\hat{s}$;

$NumStates$ is the number of states;

$Prob(inclusion \mid x_i, state = s)$ is the probability of inclusion for a student with a vector of control variables, x_i, living in state s;

$Weight_i$ = sampling weight for student *i*; and

i indexes students and s indexes states.

The measure for comparison across states is the difference between a state's actual initial-period inclusion rate and this predicted inclusion rate.

$$StateLevelActual_{\hat{s}} = \frac{1}{N_{\hat{s}}} \sum_{i=1}^{N_{\hat{s}}} Included_i \cdot Weight_i$$

$$DistAbovePredicted_{\hat{s}} = StateLevelActual_{\hat{s}} - StateLevelPredicted_{\hat{s}}$$

SUMMARY AND COMPARISON OF APPROACHES

The two approaches discussed above were developed out of conversations between NCES and AIR staff with the members of the NAEP Validity Studies panel and the Education Information Management Advisory Consortium (EIMAC). We present them with no preference for one over the other. The nation-based approach for measuring change has the advantage of relying on a single regression estimated using all states in the initial period. The large number of observations used in this regression allows us to estimate the relationship between inclusion and controls with a greater level of detail. Because it uses data from all states, we are able to create interaction terms between type of disability, severity level, and grade level of instruction variables for more detail in distinguishing student characteristics. There are not enough observations in a single state to accurately estimate all those interaction terms jointly and so only type of disability and severity level are crossed in the state-specific model.

Additionally, under the nation-based approach, a subset of the results used to determine the measure of change can be used to create the starting point measure. In contrast, the state-specific approach to measuring change requires 51 separate state-level regressions for the change measure plus an additional model for the measure of differences between states. Because the 51 regressions for the change measure are at the state level, we are able to cross type of student disability only with severity level, leaving grade level of instruction to be included as its own set of indicators. The advantage of the state-specific approach for measuring change, however, is that it eliminates any potential bias resulting from the subjective interpretation of the SD Questionnaire that is correlated to the state in which a student is tested.[12] Under the state-specific approach, a different relationship between inclusion on NAEP and student characteristics is estimated for each state, which allows change in each state to be measured by its own implicit standards as set in the initial period. Tables 3 and 4 summarize the nation-based and state-specific approaches to measuring change and methods for measuring the starting point.

Table 3. Summary and comparison of nation-based and state-specific measures of change

	Nation-based approach	**State-specific approach**
Methodology	Partitioning technique derived from Oaxaca-Blinder decomposition technique	
Result	Nation-based measure of change	State-specific measure of change
Population for fixing relationship between inclusion and controls	National NAEP SD sample (except ELLs) for initial period (2005)	For each state/jurisdiction: that state/jurisdiction's SD sample (except ELLs) for the initial period (2005)
Controls	5 disability-type indicators X 4 severity-level indicators X 4 grade level of instruction indicators, indicator of received an accommodation on state assessment that is not allowed on NAEP	5 disability-type indicators X 4 severity-level indicators, 4 grade level of instruction indicators
Benefit of approach	More interactions between the control variables	Separate relationship estimated for each state, thus circumventing potential bias due to differential interpretation of SD questionnaire across states

12 A further discussion of this point is provided in the section on caveats and cautions in interpretation.

Table 4. Summary and comparison of nation-based and state-specific measures of starting point

	Nation-based approach	State-specific approach
Result	Nation-based measure of starting point	State-specific measure of starting point
Population for fixing relationship between inclusion and controls	National NAEP SD sample (except ELLs) for initial period (2005)	National NAEP SD sample (except ELLs) for initial period (2005)
Include state fixed effects	No	Yes
Benefit of approach	Uses same regression and results as in the nation-based approach's measure of change	Including state fixed effects explicitly estimates differences between states

THE ROLE OF ACCOMMODATIONS AND STATE POLICIES

In developing these measures, we paid particular attention to the role that accommodations and state policies on inclusion on state assessments play in the inclusion of SDs in the NAEP assessment. Whether or not an SD can participate in the NAEP assessment is determined by the child's school and supported by information in the SD Background Questionnaire. Changes in NAEP inclusion rates are, therefore, likely related to the testing policies of assessment programs in a student's state because this local decision making regarding a student's participation in NAEP is likely to be heavily influenced by the rules for the participation of SDs on state assessments. Theoretically, states can include a given student without accommodation, accommodate the student (i.e., include the student with an accommodation), or not include the student.

The concern over the role of accommodations and state policies in our measures of change over time and differences among states has several facets. First, if state policies on inclusion are likely to influence how the SD Questionnaire respondent recommends a student be treated on NAEP, should those state policies be controlled for in our measure of change and/or our starting point measure? Potential information that could be used includes whether the student was excluded or included with or without accommodation and what type of accommodation he or she received on the state assessment. For our analysis, we include information on accommodations not allowed on NAEP that are provided for the student on state assessments in models that make comparisons among states. This includes the single regression model used in the nation-based approach and the regression model used for the measure of starting point, but not the measure of change over time, in the state-specific approach.

Including an indicator for *receiving an accommodation on the state assessment not allowed on NAEP* in the regression means that students who receive such an accommodation will have an adjusted probability of inclusion. In other words, with other characteristics held constant, students receiving such an accommodation are compared in inclusion treatment with other students receiving such an accommodation and not with others. This is similar to the way that students with a *specific learning disability* are compared with other students with a *specific learning disability* and not with students with *mental retardation*.

It was decided that it is unfair to states that are more accommodating than NAEP to be compared similarly with other less accommodating states for determining the predicted inclusion rate for each state. Including in the regression model an indicator for *receiving an accommodation on the state assessment not allowed on NAEP* addresses this. Further, it was also decided not to include an indicator for whether the student was excluded from a state assessment, a factor that we had considered using in the models. Such a measure would likewise set students excluded on state exams separate from other students. The purpose of this study is to gauge improvement in inclusion. Using an indicator for students given an accommodation on the state assessment that is not allowed on NAEP, in essence, would reward a state for extra efforts at accommodation by setting a separate standard of inclusion. Using an indicator for students excluded on state assessments would set a separate standard for states that are less accommodating and is hence omitted.

A control variable for students receiving an accommodation on the state assessment that is not allowed on NAEP is appropriate for any model that is estimated using more than one state's data. This control variable is included in the one regression model estimated for the nation-based approach as well as in the second regression model estimated in the state-specific approach that is used for calculating the starting point measure. The measure of change over time in the state-specific approach compares states with themselves over time, so no information about the treatment of students on state assessments is included among the control variables.

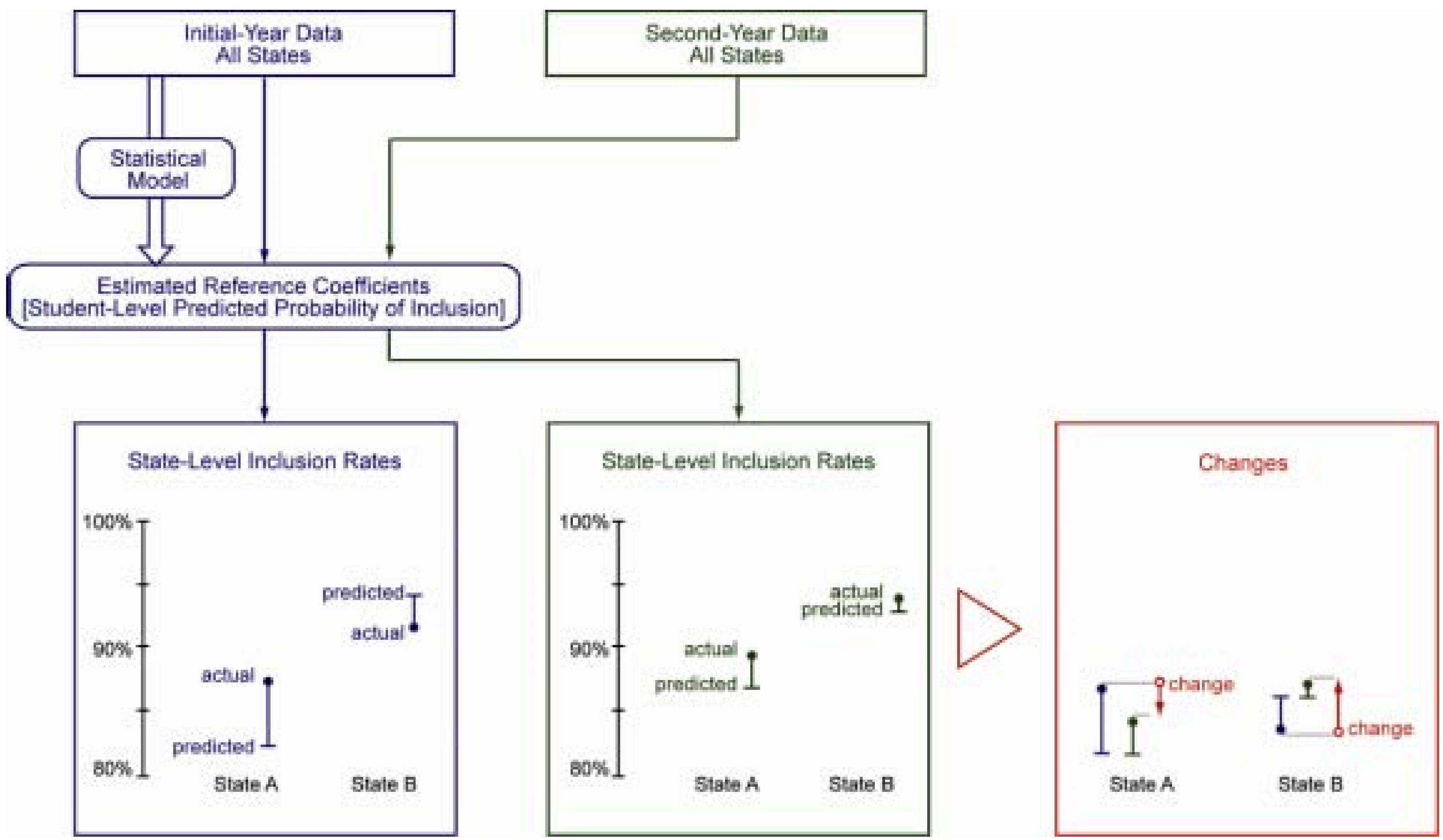
The decision to omit the control for *accommodation on the state assessment not allowed on NAEP* from the state-specific model means the following: changes in accommodations policy on state assessments that lead to changes in inclusion on NAEP *are* counted as part of that change measure. Because they are not controlled for, any effect they have is captured in the change measure. Therefore, if a state begins to allow an accommodation on its state assessments that is not allowed on NAEP and this leads to lower inclusion rates on NAEP for some students, this will show up as reduced measured change in NAEP inclusion.

ILLUSTRATIONS OF THE APPROACHES

Nation-Based Approach

To better understand how the approaches for measuring change work and how they are different, we present two graphical displays and hypothetical examples. As illustrated in figure 2, the nation-based approach uses initial-year data to estimate the relationship between student characteristics and the probability of inclusion. This estimated model provides the reference coefficients used for all states which, in turn, provide a predicted probability of inclusion for each student. These predicted probabilities of inclusion are then aggregated to provide a predicted inclusion rate for a given state in a given year. Change, as illustrated in the figure, is measured as change in the distance between the state's actual and predicted inclusion rates from the first period to the second.

Figure 2. Illustration of nation-based approach for measuring change



Consider a hypothetical situation in which there are only two types of students with disabilities in the grade 4 mathematics NAEP assessment:

Type 1:

- The student's disability is classified as *having a specific learning disability.*
- The disability is classified as *moderate.*
- The student is receiving a level of instruction in mathematics that is the *same* as the grade the student is in (grade 4).
- The student is not receiving an accommodation on the state assessment that is not allowed on NAEP.

Type 2:

- The student's disability is classified as *having emotional disturbance.*
- The disability is classified as *severe.*
- The student is receiving a level of instruction in mathematics that is *two grades below* the grade the student is in (grade 4).
- The student is receiving an accommodation on the state assessment that is not allowed on NAEP.

In this hypothetical situation, there are only two states, State A and State B. The distribution of SDs between the two types described above are given in the first two rows of table 5. The distribution of SDs is different across states and across time periods. State A has a higher proportion of students of type 2 in both years, but the proportion of type 2 students declines for each state in the second period.

Table 5. Example of nation-based approach for measuring change

		Initial period		Second period		
		State A	**State B**	**State A**	**State B**	**Change**
Distribution of SDs (percentage in type)	Type 1	65.00	90.00	75.00	97.00	
	Type 2	35.00	10.00	25.00	3.00	
Student-type predicted probabilities (set by one regression using all states' initial year data)	Type 1	.95	.95	.95	.95	
	Type 2	.60	.60	.60	.60	
State-level predicted inclusion rate		82.75	91.50	86.25	93.95	
State-level actual inclusion rate		89.00	91.00	92.00	95.00	
Distance above predicted	State A	6.25		5.75		–0.50
-	State B		–0.50		1.05	1.55

In the nation-based approach, a logistic regression using all observations from all states estimates the relationship between SD characteristics and the probability of inclusion. Hypothetical estimated coefficients for this example are as follows:

$\hat{\alpha}$	= 2.00	Intercept coefficient
$\hat{\beta}_{1,2,1}$	= 0.94	Coefficient for students with specific learning disability, moderate, same grade level of instruction
$\hat{\beta}_{4,3,3}$	= -0.40	Coefficient for students with emotional disturbance, severe, two grades behind in instruction
$\hat{\beta}_{nna}$	= -1.20	Coefficient for students receiving an accommodation on the state assessment that is not allowed on NAEP

The resulting predicted probability for each combination of student characteristics is the student-level predicted probability: the probability that a student with that given set of characteristics is included on NAEP. They are calculated by first obtaining the linear combination of the coefficients:

Linear combination of coefficients for students type 1:

$$\hat{\alpha} + \hat{\beta}_{1,2,1} = 2.00 + 0.94 = 2.94$$

Linear combination of coefficients for students type 2:

$$\hat{\alpha} + \hat{\beta}_{4,3,3} + \hat{\beta}_{nna} = 2.00 - 0.40 - 1.20 = 0.40$$

Second, we transform that linear combination of coefficients to the probability scale by means of the logistic function, $\Lambda(z) = \frac{e^z}{1+e^z}$:

Predicted probability for students type 1: $\frac{e^{2.94}}{1+e^{2.94}} = 0.95$

Predicted probability for students type 2: $\frac{e^{0.40}}{1+e^{0.40}} = 0.60$

The student-level predicted probabilities for the two types in this hypothetical example are provided in rows 3 and 4 in table 5. These predicted probabilities are the same across states and across time for all students with the same characteristics.

The state-level predicted inclusion rates are an aggregation of student-level predicted probabilities according to the distribution of the types of SDs in the state. Because of the different distributions of students, the state-level predicted inclusion rates vary across states and across time. The calculation of state-level predicted inclusion rates is straightforward in this simplified example: for State A in the initial period, 65 percent of the students are type 1 and are expected to be included at a rate of 95 percent, whereas the remaining 35 percent of students are type 2 and are expected to be included at a rate of 60 percent. The state-level predicted inclusion rate can thus be seen as a weighted average of those student-type predicted probabilities where the weights are the proportions of students in each type. For State A in the initial time period, (65% × .95) + (35% × .60) = 82.75 percent. Across states in the initial period, because State B has a greater proportion of students who are easier to include, type 1, than State A, State B's state-level predicted inclusion rate is higher than that of State A. Because in State A the proportion of students of type 1 is higher in the second period than in the initial period, State A's state-level predicted inclusion rate is higher in the second period than in the first period.

Change is measured by comparing the *distance above the predicted inclusion rate* in each period. Actual (i.e., unadjusted) inclusion rates for our example are provided in table 5. The last two rows of table 5 contain the distance above the predicted inclusion rate measures for State A and State B. State A was 6.25 percentage points above its predicted inclusion rate in the initial period and 5.75 percentage points above its predicted inclusion rate in the second period, for a change of –0.50, as reported in the last column. This means that State A was relatively *less* inclusive in the second period by our measure of 0.5 percentage point. State B, however, increased its inclusion relative to its predicted inclusion rate by 1.55 percentage points.

Once we have the change measures, it is important to put them in context. That context is a comparison of the relative inclusiveness of states in the initial period. In the nation-based approach, this is comparing both states' *distance above the benchmark* in the initial period. In the initial period, State A, at 6.25 percentage points above its benchmark, is relatively more inclusive than State B, which was 0.50 percentage point *below* its benchmark. Given this context, it is not surprising to see State B improve and State A not improve.

State-Specific Approach

Building on the example for the nation-based approach, we look at an example for the state-specific approach in which the regression model used to fix the relationship between student characteristics and the probability of inclusion is estimated separately for each state using the initial period's data, as illustrated in figure 3. Estimation of the statistical model is done separately for each state resulting in a separate set of reference coefficients for each state. Those coefficients are then used with their respective state's second period data to provide a predicted probability of inclusion for each student. The predicted probabilities are aggregated within the state to obtain a predicted inclusion rate for that state for the second period. Change for a state, under the state-specific approach, is the difference between the states actual inclusion rate and that predicted by the model.

Hypothetical results from these regressions are as follows:

State A

$\hat{\alpha}^A$ = 4.70 Intercept coefficient

$\hat{\beta}_{1,2}^A$ = -0.10 Coefficient for students with specific learning disability, moderate

$\hat{\beta}_{4,3}^A$ = -1.30 Coefficient for students with emotional disturbance, severe

$\hat{\gamma}_3^A$ = -2.78 Coefficient for students two grades behind in instruction

State B

$\hat{\alpha}^B$ = 2.75 Intercept coefficient

$\hat{\beta}_{1,2}^B$ = -0.31 Coefficient for students with specific learning disability, moderate

$\hat{\beta}_{4,3}^B$ = -0.75 Coefficient for students with emotional disturbance, severe

$\hat{\gamma}_3^B$ = -2.28 Coefficient for students two grades behind in instruction

For a given set of student characteristics, therefore, each state will have its own student-level predicted probability set by the initial period, as given for students of type 1 and type 2 in the first two rows of table 6. These are also calculated by first obtaining the linear combination of the coefficients and then transforming them to the probability scale using the logistic function:

State A

Linear combination of coefficients for students type 1[13]:

$\hat{\alpha}^{A} + \hat{\beta}^{A}_{1,2}$ = 4.70 – 0.10 = 4.60

Linear combination of coefficients for students type 2:

$\hat{\alpha}^{A} + \hat{\beta}^{A}_{4,3} + \hat{\gamma}^{A}_{3}$ = 4.70 – 1.30 – 2.78 = 0.62

State B

Linear combination of coefficients for students type 1:

$\hat{\alpha}^{B} + \hat{\beta}^{B}_{1,2}$ = 2.75 – 0.31 = 2.44

Linear combination of coefficients for students type 2:

$\hat{\alpha}^{B} + \hat{\beta}^{B}_{4,3} + \hat{\gamma}^{B}_{3}$ = 2.75 – 0.75 – 2.28 = –0.28

State A

Predicted probability for students type 1: $\frac{e^{4.60}}{1+e^{4.60}}$ = 0.99

Predicted probability for students type 2: $\frac{e^{0.62}}{1+e^{0.62}}$ = 0.65

State B

Predicted probability for students type 1: $\frac{e^{2.44}}{1+e^{2.44}}$ = 0.92

Predicted probability for students type 2: $\frac{e^{-0.28}}{1+e^{-0.28}}$ = 0.43

13 There is no coefficient for the grade-level of instruction for students of type A because the category for those receiving a grade-level of instruction at or above the grade level is the omitted, or reference, category.

Figure 3. Illustration of state-specific approach for measuring change

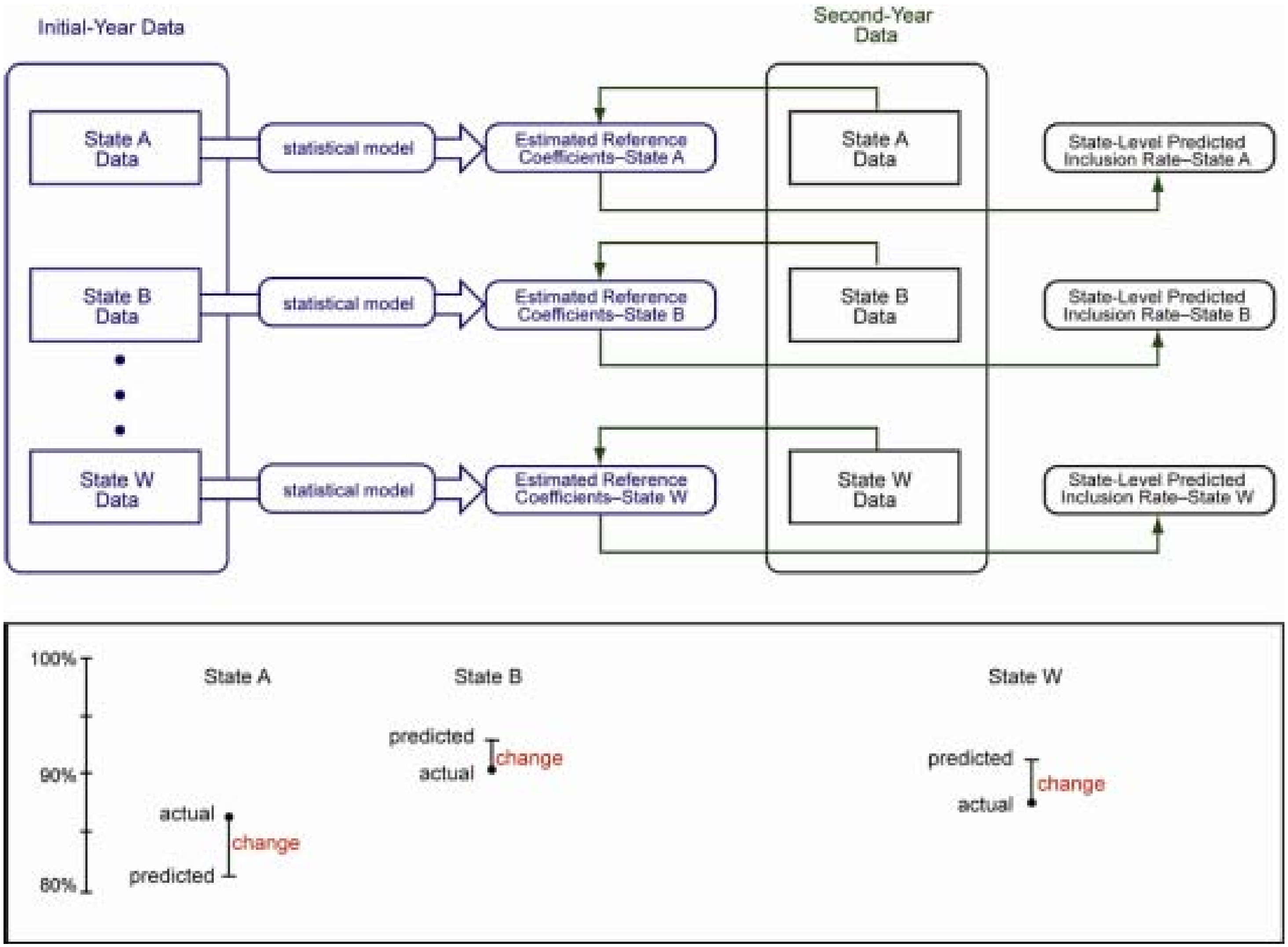


Table 6. Example of state-specific approach for measuring change

		State A	State B
Student-type predicted probabilities (set by separate regressions for each state using initial year data)	Type 1	.99	.92
	Type 2	.65	.43
Distribution of SDs (percentage in type)	Type 1	75.00	97.00
Second-year data	Type 2	25.00	3.00
State-level predicted inclusion rate for second year		90.50	90.53
State-level actual inclusion rates for second year		90.00	95.00
Change		–0.50	4.47

The third and fourth rows of table 6 have the distribution of students for State A and State B in the second period. The state-level predicted inclusion rates are a weighted average of the student-type benchmarks using the proportion of students in each type as a weight. In our example, State A's state-level benchmark is (99% × .75) + (65% × .25) = 90.5 percent. This state-level predicted inclusion rate is the inclusion rate we would expect that state to have because of the rates by which it included different types of students in the first period and on the proportions of students in each type in the second period.

Comparing the actual (unadjusted) second-period inclusion rates with the state-level predicted inclusion rate gives the measure of change. State A is predicted to have a 90.5 percent inclusion rate on the basis of its student-level predicted probabilities set in the initial period. State A's actual inclusion rate in the second period is 90 percent, meaning that it was less inclusive in the second period than in the initial period. For State B, the actual inclusion rate in the second period is nearly 5 percentage points higher than its predicted inclusion rate, indicating that it is more inclusive in the second period.

Again, the change measures need to be put into context. In the state-based approach, this requires a separate regression. The regression is distinct from the regression used in the nation-based approach but is similar enough that in this simple exercise they produce the same results. Hence, we simply refer back to table 5 where State A had a higher inclusion rate adjusted for differences in SD population in the initial period. As in the nation-based approach, we again conclude that although State B increased its relative inclusion of students, it also started out relatively less inclusive; so, it is not surprising that it had a larger increase in the change measure.

STARTING POINT VS. CHANGE

Under both approaches, we can compare states with one another in the initial period using the *starting point* measure. The starting point measure provides a context for the measure of change over time, which is the focus of this study. Because it is not easy to evaluate these two measures at the same time, we simplified and combined them by categorizing the two-dimensional display space: starting point *versus* change. For the starting point, the measure is given a quartile rank: all the states' starting point measures (distance above the benchmark measures for the initial period) are ordered and partitioned into quartiles. Each state is subsequently assigned a number from 1 to 4, according to which quartile it is in, with 4 being the highest quartile (the top 25 percent of starting point measures) and 1 the lowest (the bottom 25 percent of starting point measures).

Change is of most interest for this study; we summarize the change measure by whether the change is statistically different from zero and give its direction if it is statistically significant.[14] For change, we assign the state a score of 1 if the change measure is both positive and statistically significant (the state is more inclusive of SDs in 2007 than in 2005), 0 if the change measure is not statistically different from zero, and −1 if the change measure is both negative and statistically significant (the state is less inclusive of SDs in 2007 than in 2005). We then assign each state a *composite index score*, which uses these two scores as coordinates. This divides the *starting point vs. change* space into 12 bins, shown in figure 4. For each subject and grade assessment in NAEP, every state falls into one of these bins.

Figure 4. Composite index score by quartile of starting point score and statistical significance of change score

Starting point quartile				
4 more inclusive	(4, –1)	(4, 0)	(4, 1)	
3	(3, –1)	(3, 0)	(3, 1)	
2	(2, –1)	(2, 0)	(2, 1)	
1 less inclusive	(1, –1)	(1, 0)	(1, 1)	
	–1 negative and significant	**0 no significant change**	**1 positive and significant**	**Change**

This partitioning of the space simplifies the understanding of results by focusing on statistically significant change in inclusion rates while providing a context for understanding that change. A priori, we expect to find states making positive and significant change to be located lower on the scale of inclusiveness. A more nuanced evaluation can be performed by looking directly at the values of the measures, but this captures the relative essence of those results to facilitate their understanding. In the results tables, we provide the values of the measures as well as this simplified composite index score.

14 All tests were conducted at the 95 percent confidence level using simple *t*-tests. Estimation of standard errors is described in appendix A.