Measuring the Power of Learning.®

Research Report
ETS RR–17-17

Long-Term Impact of Valid Case Criterion on Capturing Population-Level Growth Under Item Response Theory Equating

Weiling Deng

Lora Monfils

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Long-Term Impact of Valid Case Criterion on Capturing Population-Level Growth Under Item Response Theory Equating

Weiling Deng & Lora Monfils

Educational Testing Service, Princeton, NJ

Using simulated data, this study examined the impact of different levels of stringency of the valid case inclusion criterion on item response theory (IRT)-based true score equating over 5 years in the context of K–12 assessment when growth in student achievement is expected. Findings indicate that the use of the most stringent inclusion criterion generally yielded the most accurate results when overall root mean square error (RMSE) and bias were considered under both zero-growth and growth conditions, for both one-parameter logistic (1PL) and three-parameter logistic (3PL) IRT models, and for both fixed common item parameter (FCIP) and test characteristic curve (TCC) scaling methods. The positive impact of applying the most stringent valid case inclusion criterion was more salient with the 3PL model, under which greater classification accuracy was observed.

Keywords IRT; equating; valid case (inclusion) criterion; growth

doi:10.1002/ets2.12144

An important requirement of equating is population invariance (Dorans & Holland, 2000). To date, a number of studies have looked at population invariance across subgroups and methods (Liu & Holland, 2008; Schmitt, Cook, Dorans, & Eignor, 1988; von Davier & Wilson, 2008; Yang, 2004; Yang & Gao, 2008; Yi, Harris, & Gao, 2008), but research on the implications of population invariance, or lack thereof, for capturing change or growth in student achievement at the population level[1] is limited. In contexts where standards-based assessments are used to hold schools and jurisdictions accountable for demonstrating progress in improving student achievement, such as the No Child Left Behind Act (2001) (NCLB), yet reporting timelines or other policy considerations may influence the choice of equating sample, the degree to which different valid case criteria influence the accuracy of reported results over multiple years is unclear.

Item response theory (IRT) true score equating methods, distinguished by the scaling method used to place item parameters onto a common theta metric, are often used for K–12 statewide assessments with the nonequivalent groups with an anchor test (NEAT) data collection design. The assumptions required in applying this methodology in operational settings are fivefold (von Davier & Wilson, 2007): (a) Two random independent samples drawn from two populations take two different test forms, which share a common anchor test; (b) the assumptions of unidimensionality, local independence, and monotonicity hold for the anchor test and the total tests—they all measure the same construct; (c) the anchor test item parameters are population invariant (thus sample invariant) if the model fits the data for each of the two populations—due to scale indeterminacy, a linear transformation is needed to place the two sets of parameters on the same scale; (d) IRT equating assumes that there are no omitted responses; and (e) the relationship between true scores generalizes to the observed ones.

Robustness of equating to differences in samples used to obtain the equating functions has been studied for IRT equating methods (Brennan, 2008; Kolen, 2004). Overall, IRT equating has been shown to be population invariant (Eignor, Stocking, & Cook, 1990; Stocking, Eignor, & Cook, 1988; von Davier & Wilson, 2008). In addition, Skaggs and Lissitz (1986) showed that equating based on the three-parameter logistic (3PL) model was more robust than that based on the one-parameter logistic (1PL) or Rasch model.[2]

Various studies have compared IRT equating results based on different scale transformation methods, and it seems that the best method for scaling depends on the context under which tests are equated. Concurrent calibration (CC) is more efficient than other equating procedures when the model holds for the data (Beguin, 2002, 2009; Kim & Cohen,

Corresponding author: W. Deng, E-mail: wdeng@ets.org

2002; Kolen & Brennan, 2004). Overall, however, Stocking and Lord's (1983) test characteristic curve (TCC) method may provide better results due to its robustness in the presence of a modest amount of error (Kolen & Brennan, 2004).

Several studies have investigated the long-term sustainability of various IRT equating methods in capturing population-level growth over time. Using empirical data from a criterion-referenced test, Jodoin, Keller, and Swaminathan (2003) compared three IRT equating methods—fixed common item parameter (FCIP), mean-sigma (MS), and CC—and two ability estimates (the maximum likelihood ability estimate [MLE] and the expected a posteriori ability estimate [EAP]) with respect to recovering growth for a mixed-format test. They found that linear equating indicated less growth than the FCIP or CC methods and that the MLE estimates reflected larger growth than the EAP estimates.

Using data simulated with empirical parameters from a statewide testing program, Baldwin and Baldwin (2007) studied the effect of anchor test length on the recovery of item parameters and increase in ability across four administrations in a mixed-format test for five IRT equating methods—FCIP, TCC, MS, Haebara's (1980) item characteristic curve (ICC) method, and mean and mean (MM). They found that all the methods performed similarly when ability distributions remained the same, but this was not the case when ability distributions changed from year to year, whether the increase was uniform or skewed.

Using simulated data, Keller and Keller (2011) compared FCIP, TCC, ICC, MS, and MM scaling over six administrations with respect to bias and root mean square error (RMSE) of theta estimates under different types of shifts in the ability distribution. They found that FCIP provided the most accurate long-term results when both mean and skewness changed in the ability distribution. One limitation of their study is that the same operational and internal anchor test form was used for the six consecutive administrations "to eliminate the confounding of test length and composition as sources of error in the equating function" (Keller & Keller, 2011, p. 365). A logical extension to inform practice would be to compare the various scaling methods under realistic data collection designs.

According to von Davier and Wilson (2007), IRT equating rests on the assumption that there are no omitted responses. In reality, missing data exist in almost any assessment. In addition, missing responses do not occur randomly, and the rates of omission vary as a function of ability, personality characteristics, and test-taking strategies (Ayala, Plake, & Impara, 2001; DeMars, 2003). Different ways of treating missing responses will cause differences in item parameter and ability estimates. Typical treatments include incorrect, not presented, fractionally correct, and the two-stage method where non-responses are first coded as *not presented* for item calibration and then *incorrect* for ability estimation (Ayala et al., 2001; Finch, 2008; Ludlow & O'Leary, 1999; Shin, 2009).

Using simulated data, Deng and Monfils (2011) examined the impact of different levels of stringency in valid case criteria on Rasch IRT equating over five administrations of a K–12 assessment under three conditions of growth in student achievement. They found that applying the most stringent valid case criterion generally yielded the most accurate equating results; also, under conditions studied, FCIP outperformed the MM method. Their study used Winsteps for calibration and equating, applying the default setting of coding missing responses as incorrect. The purpose of this study is to expand the previous study to investigate the long-term impact of different stringency levels of the valid case criterion on capturing population-level growth in academic achievement under two IRT models, 1PL and 3PL, and two popular equating methods, FCIP and TCC.

Method

Design

The simulation study was designed to reflect a wide range of what might be seen in practice across various testing programs and conditions of assessment. Student-level item response data were simulated for five successive yearly administrations of a two-section[3] multiple-choice test.[4] Three levels of stringency in valid case criteria were studied under both 1PL and 3PL IRT models. The most stringent level required that only those examinees who had attempted all test items be included in calibration. In the medium stringency condition, all those who had attempted 10 or more items in each section constituted the calibration sample. The least stringent valid case criterion included all examinees who had attempted at least one item in any section. For changes in the ability distribution, three conditions of uniform growth (i.e., same amount of growth across the ability scale) were considered. Under zero or no growth, a standard normal ability distribution was used each year. Small growth was defined by a yearly increase in average ability of .10 in the theta metric and large growth by a yearly increase of .25. Note that all three growth conditions started with ability distributed as $N(0,1)$ in Year 1. Equating for all

Table 1 Simulation Design of the Study

Factors	No. of levels	Description
Stringency of valid case criterion	3	Most stringent, medium stringency, least stringent
Item response theory model	2	1PL, 3PL
Equating method	2	FLIP, TCC
Growth condition and ability distributions for years 1 – 5	3	Zero growth: $N(0,1)$, $N(0,1)$, $N(0,1)$, $N(0,1)$, $N(0,1)$; Small growth: $N(0,1)$, $N(.10,1)$, $N(.20.1)$, $N(.30.1)$, $N(.40,1)$; Large growth: $N(0,1)$, $N(.25,1)$, $N(.50,1)$, $N(.75,1)$, $N(1.00,1)$

Note. 1PL = one-parameter logistic method; 3PL = three-parameter logistic method; FCIP = fixed common item parameter method; TCC = test characteristic curve method.

Table 2 Summary Statistics of "True" Item Parameters for Total Test and Anchor Test

	Year 1			Year 2			Year 3			Year 4			Year 5		
Total test[a]	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
Mean	1.01	−.21	.20	1.01	−.25	.21	.99	−.07	.19	.99	−.23	.22	1.00	−.14	.23
SD	.16	1.51	.09	.15	1.61	.09	.13	1.58	.09	.13	1.49	.09	.13	1.52	.08
Min.	.65	−2.67	.03	.65	−2.78	.03	.71	−2.78	.00	.71	−2.70	.05	.61	−2.66	.00
Max.	1.46	2.62	.42	1.46	2.61	.37	1.31	2.74	.36	1.24	2.74	.46	1.24	2.41	.46

	Anchor 12[c]			Anchor 23			Anchor 34			Anchor 45		
Anchor test[b]	a	b	c	a	b	c	a	b	c	a	b	c
Mean	1.04	−.18	.20	1.00	−.47	.19	.97	−.12	.20	1.00	−.70	.26
SD	.20	1.47	.09	.12	1.78	.10	.14	1.39	.08	.13	1.47	.09
Min.	.65	−2.52	.08	.80	−2.78	.03	.71	−2.13	.05	.77	−2.66	.08
Max.	1.46	2.57	.37	1.26	2.61	.36	1.20	2.74	.32	1.24	2.36	.46

Note. The same b-values were used for both one-parameter and three-parameter logistic models.
[a] $N = 60$. [b] $N = 20$. [c] Anchor 12 is the anchor set composed of items common to years 1 and 2 forms.

conditions was based on the NEAT design with an internal anchor. Table 1 lists the simulation conditions of this study. The four factors were fully crossed, yielding a total of 36 conditions. One hundred replications were run for each condition.

Data Generation

In this study, the simulated test form included 60 multiple-choice items with two identical sections of 30 items each. Based on realistic item parameters, five parallel test forms were generated using WinGen (Han, 2012), one for each year of administration. These forms had similar a, b, and c parameters as well as a wide range of item difficulties. Item parameters were first generated for the 3PL model, and the same generated b-values were used for the 1PL model. Forms in adjacent years shared 20 items in common, that is, anchor items. Within each form, items at Positions 2, 5, 8, 11, … , 59 constituted the anchor between the preceding form and current form, and items at Positions 3, 6, 9, 12, … , 60 constituted the anchor between the current and the succeeding form. Table 2 shows the statistical characteristics of the total test forms and the associated anchor sets.

The generation of examinee response data was completed in several steps. First, for each year and growth condition, the true ability of each of the 50,000 individuals composing the population was generated from the corresponding distribution listed in Table 1. Next, complete response strings were generated for the 50,000 simulees in each year and for each growth condition. For each simulee, the probability of a correct response to each item was computed as a function of true ability and a, b, and c parameters in the corresponding year's form, according to the 3PL IRT model. The examinee's observed response to each item was obtained by drawing a random uniform number between 0 and 1. A score of 1 was assigned if the probability of a correct response to the item was greater than the random number, and a score of 0 was assigned otherwise. This resulted in 15 population data sets with complete response strings.

Then, following methods similar to those employed by Stocking et al. (1988), DeMars (2003), Finch (2008), and Ayala et al. (2001), missing responses were generated in the response strings of a subset of simulees in each population to reflect

Table 3 Rates Used to Generate Intentional Omits

	Item response	
Ability group	Correct	Incorrect
High	.08	.20
Medium	.16	.24
Low	.38	.42

Table 4 Rates Used to Generate Missing Responses Due to Speededness

Ability group	Omit rates for items 1 – 8 …
High	.00, .00, .00, .00, .00, .01, .02, .03, …
Medium	.06, .07, .08, .09, .10, .11, .12, .13, …
Low	.16, .17, .18, .19, .20, .21, .22, .23, …

the level and pattern of missingness typically observed in large-scale K – 12 statewide assessments. Approximately 80% of the simulees had no missing responses. Following Ayala et al. (2001) and Shin (2009), who noted that examinees are more likely to omit an item if they are not sure of the answer, approximately 15% of the simulees omitted a varying number of items throughout the test: these are intentional omits. Missing responses due to speededness were generated for approximately 5% of the simulees. Overall, the rate of missing responses was approximately 6% in each population.

For both intentional omits and omits due to speededness, missingness was not completely random, and consideration was made so that lower ability simulees were more likely to miss than higher ability simulees. Examinees were ranked on true ability and assigned to one of three ability groups: low, medium, or high. Examinees within each ability group were then randomly selected for generation of missing item responses of either type with probability rates detailed in the next two paragraphs. Note that starting with rates in the previously cited research, various rates were explored and evaluated considering the goals of current study.

The rates used to generate intentional omits were based on examinee ability group and the correctness of the item response, as shown in Table 3. Higher ability examinees would have lower omit rates, and the probability to omit an item would be lower if in the original data, the simulee responded correctly to that item. For example, the intentional omit rates for high-ability simulees were .08 for items with correct responses and .20 for items with incorrect responses. Corresponding rates for the low-ability group were .38 and .42, respectively. To generate intentional omits in a given examinee's response string, a random uniform number between 0 and 1 was generated for each item, and if the resulting value was lower than the assigned omit rate for that examinee's ability level and item score, the item response was changed to an omit; otherwise, the response was not changed.

The rates used to generate missing responses due to speededness were based on examinee ability group and item position. The high-ability group had no omits due to speededness for Items 1 – 5; the omit rate for Item 6 was .01, which increased for each item thereafter within the section in increments of .01. For the medium-ability group, omits due to speededness started with a rate of .06 for Item 1 and increased for each item in the section in increments of .01. For the low-ability group, the omit rate started with .16 for Item 1, which increased for each item in the section in increments of .01 (see Table 4).

Taking into account both intentional omits and omits due to speededness, the resulting degree of missing responses was similar in the population data sets of different growth conditions. For the purpose of illustration, Table 5 presents the degree of missing responses in the small-growth populations by ability group and overall.

After missingness had been created in the population data, samples of 5,000 simulees were randomly drawn from each population according to the different levels of stringency in the valid case inclusion criterion. Table 6 provides the level of missingness in samples of varying inclusion stringencies. Because there are no omits in all of the most stringent samples (i.e., the mean would be 0), those samples are not included in the table to save space.

Consistent with what was done in the previous study, in all calibrations, missing responses were treated as presented and scored as incorrect, a common practice for untimed K – 12 assessments where state policies often require that omitted responses be treated as incorrect in the case of multiple-choice items or assigned the lowest score in the case of constructed-response items.

Table 5 Average Number and Proportion of Omits in the Small-Growth Populations

Omits	Group	*N*	Mean	SD	Max.	Median	Min.
Year 1							
NOmit	Low	16,666	5.47	11.25	48.00	.00	.00
	Medium	16,667	3.14	7.47	41.00	.00	.00
	High	16,667	1.99	5.64	35.00	.00	.00
	Overall	50,000	3.533	8.573	48.000	.000	.000
PropOmit	Low	16,666	.09	.19	.80	.00	.00
	Medium	16,667	.05	.12	.68	.00	.00
	High	16,667	.03	.09	.58	.00	.00
	Overall	50,000	.059	.143	.800	.000	.000
Year 2							
NOmit	Low	16,666	5.46	11.22	48.00	.00	.00
	Medium	16,667	3.07	7.40	41.00	.00	.00
	High	16,667	1.92	5.49	38.00	.00	.00
	Overall	50,000	3.484	8.512	48.000	.000	.000
PropOmit	Low	16,666	.09	.19	.80	.00	.00
	Medium	16,667	.05	.12	.68	.00	.00
	High	16,667	.03	.09	.63	.00	.00
	Overall	50,000	.058	.142	.800	.000	.000
Year 3							
NOmit	Low	16,666	5.45	11.25	46.00	.00	.00
	Medium	16,667	3.10	7.44	43.00	.00	.00
	High	16,667	1.95	5.51	35.00	.00	.00
	Overall	50,000	3.496	8.538	46.000	.000	.000
PropOmit	Low	16,666	.09	.19	.77	.00	.00
	Medium	16,667	.05	.12	.72	.00	.00
	High	16,667	.03	.09	.58	.00	.00
	Overall	50,000	.058	.142	.767	.000	.000
Year 4							
NOmit	Low	16,666	5.59	11.37	46.00	.00	.00
	Medium	16,667	2.97	7.33	41.00	.00	.00
	High	16,667	1.93	5.52	38.00	.00	.00
	Overall	50,000	3.493	8.574	46.000	.000	.000
PropOmit	Low	16,666	.09	.19	.77	.00	.00
	Medium	16,667	.05	.12	.68	.00	.00
	High	16,667	.03	.09	.63	.00	.00
	Overall	50,000	.058	.143	.767	.000	.000
Year 5							
NOmit	Low	16,666	5.49	11.25	46.00	.00	.00
	Medium	16,667	2.96	7.21	41.00	.00	.00
	High	16,667	2.01	5.69	35.00	.00	.00
	Overall	50,000	3.487	8.512	46.000	.000	.000
PropOmit	Low	16,666	.09	.19	.77	.00	.00
	Medium	16,667	.05	.12	.68	.00	.00
	High	16,667	.03	.09	.58	.00	.00
	Overall	50,000	.058	.142	.767	.000	.000

Note. NOmit = number of omits; PropOmit = proportion of omits. Three decimal places were used for the overall results; 2 were used for each ability level.

Calibration and Equating

Calibration/equating runs were conducted separately for each of the 100 samples per condition per year. In Year 1, operational test forms were postequated back to the "bank," thus placing the Year 1 forms on the scale of the generating item parameters. In Years 2–5, operational test forms were placed on the bank scale by equating to the on-scale forms of the previous administration. This design facilitated analyses relative to the generating item and person parameters.

IRT parameter estimates for each of the 5 years of data were determined using a proprietary version of the IRT estimation program PARSCALE (Muraki & Bock, 1999). This version of PARSCALE allows for a more flexible and stable estimation of item parameters, as it employs a semi-Bayesian approach to fitting the IRT model.

Table 6 Average Number and Proportion of Omits in Samples of Varying Valid Case Inclusion Stringencies: Mean Over Replications

			Zero growth		Small growth		Large growth	
Year	Inclusion stringency	Growth condition	Mean	SD	Mean	SD	Mean	SD
Year 1	Least	NOmit	3.53	.12	3.53	.12	3.53	.12
		PropOmit	.06	.00	.06	.00	.06	.00
	Med.	NOmit	2.84	.08	2.84	.08	2.84	.08
		PropOmit	.05	.00	.05	.00	.05	.00
Year 2	Least	NOmit	3.48	.11	3.47	.11	3.47	.11
		PropOmit	.06	.00	.06	.00	.06	.00
	Med.	NOmit	2.77	.09	2.76	.08	2.76	.10
		PropOmit	.05	.00	.05	.00	.05	.00
Year 3	Least	NOmit	3.51	.12	3.49	.12	3.48	.11
		PropOmit	.06	.00	.06	.00	.06	.00
	Med.	NOmit	2.77	.11	2.76	.09	2.75	.10
		PropOmit	.05	.00	.05	.00	.05	.00
Year 4	Least	NOmit	3.52	.11	3.50	.11	3.48	.11
		PropOmit	.06	.00	.06	.00	.06	.00
	Med.	NOmit	2.74	.10	2.73	.10	2.72	.10
		PropOmit	.05	.00	.05	.00	.05	.00
Year 5	Least	NOmit	3.50	.10	3.49	.10	3.47	.10
		PropOmit	.06	.00	.06	.00	.06	.00
	Med.	NOmit	2.79	.09	2.78	.09	2.77	.09
		PropOmit	.05	.00	.05	.00	.05	.00

Note. NOmit = number of omits; PropOmit = proportion of omits.

Two different equating methods were evaluated in this study. The FCIP method does not require a separate linking step; rather, it places the non-anchor items in the new form on the operational scale during the calibration phase by fixing the anchor item parameter estimates to their previously estimated operational (or banked) values. In the TCC method, two independent calibrations take place, one for the new operational form and one for the reference form. After calibration, the new form estimates are transformed onto the reference form scale by minimizing the distance between the anchor TCCs. The TCC method was implemented using the STUIRT program (Kim & Kolen, 2004).

Evaluative Measures

Using generating item parameters, "true" scoring tables were obtained from the test characteristic curve for each administration. Theta cutscores were −1.20, 0, and 1.20, for basic, proficient, and advanced, respectively. Thetas were linearly transformed onto a typical reporting scale using additive and multiplicative transformation constants of 400 and 40. "Operational" scoring tables were obtained using on-scale parameter estimates from IRT calibration and equating. Scale scores based on the true scoring table derived from the generating parameters were the criterion against which equated scale scores were compared.

Overall Equating Accuracy

Three criteria (Harris & Crouse, 1993) were used to evaluate the overall accuracy of each equating condition for each year and across 5 years: (a) RMSE, (b) bias, and (c) standard error of estimation (SE).

RMSE provides a measure of the overall accuracy of an equating. It is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_i f_i \left[\hat{e}_Y\left(x_i\right) - e_Y\left(x_i\right)\right]^2}{\sum_i f_i}},$$

where f_i is the number of examinees with raw score x_i on the new form, $e_Y(x_i)$ is the criterion scaled score on the reference form, and $\hat{e}_Y(x_i)$ is the corresponding scaled score on the reference form as determined by any equating method (Harris & Crouse, 1993, p. 203). The overall bias relative to the criterion is defined as

$$\text{Bias} = \frac{\sum_i f_i \left[\hat{e}_Y\left(x_i\right) - e_Y\left(x_i\right)\right]}{\sum_i f_i}.$$

The bias term is part of the RMSE and can be found by decomposing the RMSE into the variance of the difference plus the squared bias (Harris & Crouse, 1993):

$$\frac{\sum_i f_i \left(d_i\right)^2}{\sum_i f_i} = \frac{\sum_i f_i \left(d_i - \bar{d}\right)^2}{\sum_i f_i} + \bar{d}^2;$$

that is,

$$\text{RMSE}^2 = \text{SE}^2 + \text{Bias}^2,$$

where d_i is $\hat{e}_Y(x_i) - e_Y(x_i)$ and $\bar{d}$ is the mean difference, that is, the bias term.

The first term on the right side of the preceding equation represents the variance of the difference between values of the equating function under evaluation and the criterion. The square root of the variance constitutes another criterion:

$$\text{SE} = \sqrt{\frac{\sum_i f_i \left[\hat{e}_Y\left(x_i\right) - e_Y\left(x_i\right) - \text{Bias}\right]^2}{\sum_i f_i}}.$$

Classification Accuracy

In the context of K–12 statewide assessment and NCLB reporting of testing results, a key accountability indicator is the percentage of students meeting a designated performance standard (i.e., PAA, or percent at or above proficient). Each examinee is classified into a specific performance category based on his or her test score to facilitate adequate yearly progress calculations. In this study, simulees' true performance categories can be known based on their true ability relative to the three designated theta cutscores that determine four performance categories. On the other hand, as in practice, examinees have equated test scores, based on which they can be classified into one of the four performance categories. Classification accuracy, defined as the total percentage of simulees correctly classified across the four performance categories when comparing the equating-based performance category against the true category, was used as an additional criterion to evaluate the impact on equating of varying stringencies in the valid case criterion.

Results

The mean RMSE over 100 replications is shown for Years 2–5 for the three calibration sample valid case criteria for different growth conditions and equating methods for the 1PL model in Figure 1 and for the 3PL model in Figure 2.

Examination of the figures indicates that with FCIP, for both 1PL and 3PL models across all growth conditions, the RMSE decreased as the level of valid case inclusion stringency increased, and equating was most accurate under the most stringent valid case criterion. The differences among the three levels of inclusion stringency were more distinct under 3PL than under 1PL. Under 3PL, the RMSE increased as the years went by. In contrast, under 1PL, the RMSE was generally stable across the years and only showed a notable increase in the large-growth condition from Years 4 to 5.

Compared to FCIP, the pattern of results is somewhat similar but less clear-cut with TCC equating. Under the 1PL model, in the zero-growth condition, when the ability distribution remained unchanged over the years, the RMSE was similar for all three stringency levels of the valid case criterion and considerably smaller than that with FCIP. However, under large growth, where average examinee ability increased by .25 each year, the RMSE was the smallest for the most

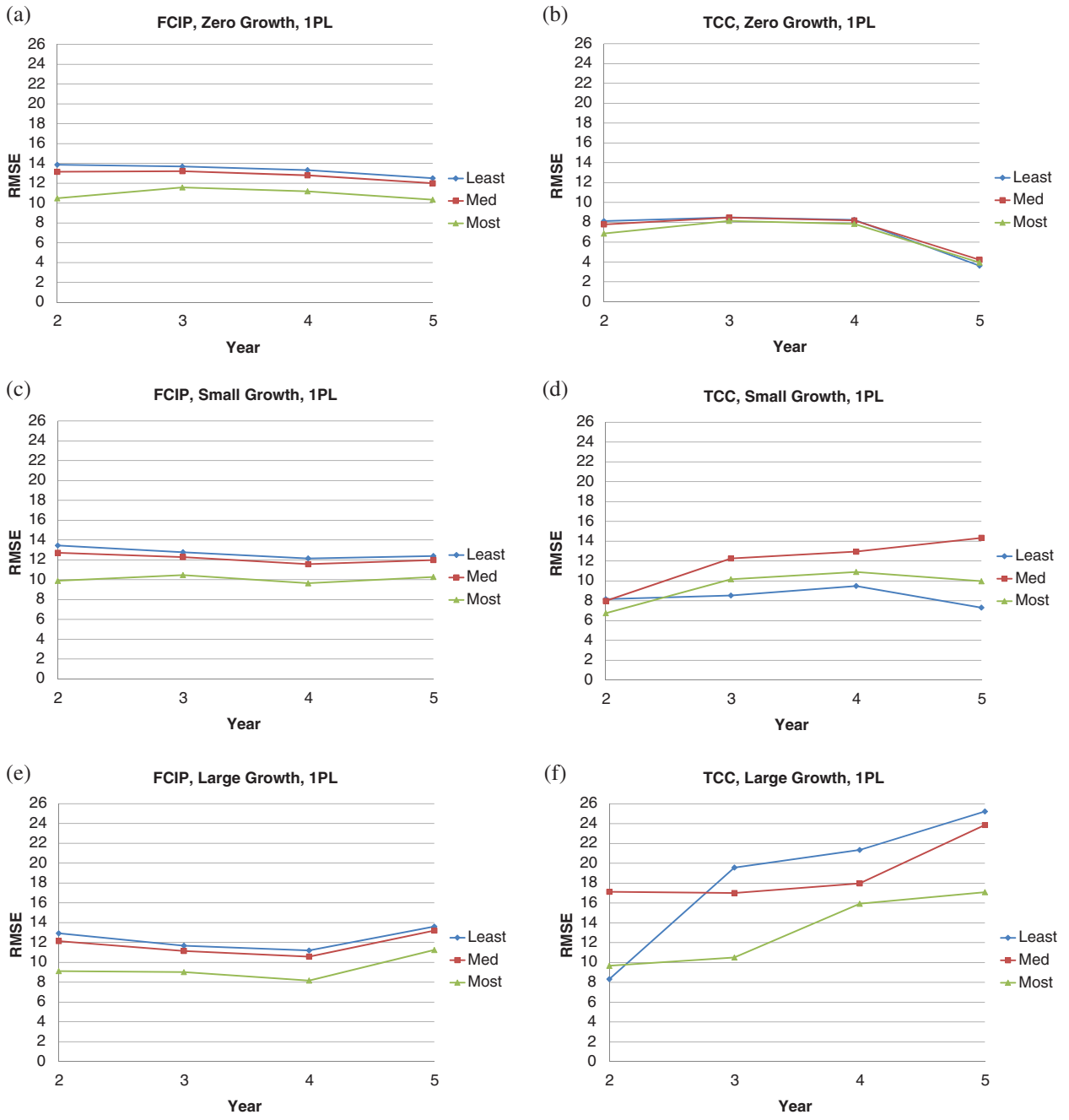

Figure 1 Mean root mean square error by year under the one-parameter logistic model.

stringent valid case criterion. Under 3PL for both the zero-growth and small-growth conditions, applying the most stringent valid case criterion resulted in by far the smallest RMSE compared to the other two valid case criteria. However, under the large-growth condition, after Year 3, the RMSE increased linearly for the most stringent sample and became the highest of the three samples in Year 5.

An examination of the mean bias in Figures 3 and 4 helps to provide some insight into the RMSE results. The general pattern of the bias results is largely similar for the two equating methods, especially under 3PL. Overall, it appears that a negative bias accumulates over administrations for both FCIP and TCC under both IPL and 3PL models. That is, student achievement became more and more underestimated as the equating chain unfolded. Under the 3PL model, Year 2 equated scores for the calibration sample with complete data had the smallest amount of bias—near zero and negative—but

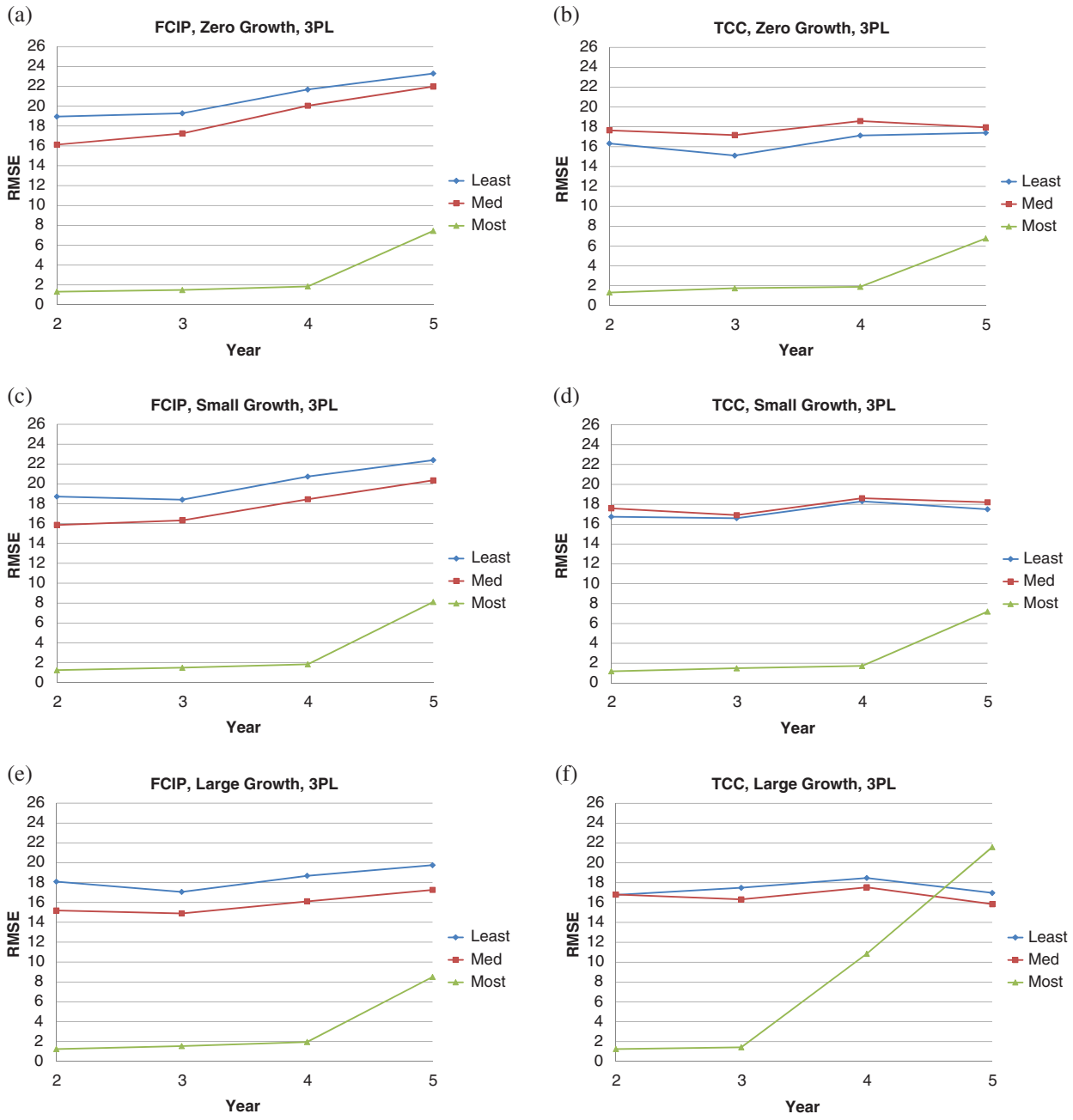

Figure 2 Mean root mean square error by year under the three-parameter logistic model.

over administrations ended up with the largest amount of negative bias in Year 5. In contrast, equated scores for the two calibration samples with varying degrees of missing responses began with some positive bias but ended up with the smallest amount of absolute bias.

Table 7 gives the average of the mean RMSEs across the equating years (i.e., Years 2 – 5). Of the three valid case criteria, the most stringent criterion was generally associated with the smallest average RMSE across growth conditions, equating methods, and IRT models. Note that TCC usually had lower average RMSE than FCIP under the zero-growth and small-growth conditions across the valid case criteria. Under large growth, though, the RMSE was mostly larger for TCC than for FCIP.

Table 8 provides the average of the mean biases across equating years. Bias was the smallest in magnitude under the most stringent valid case criterion for both IRT models and both equating methods under the no-growth and small-growth

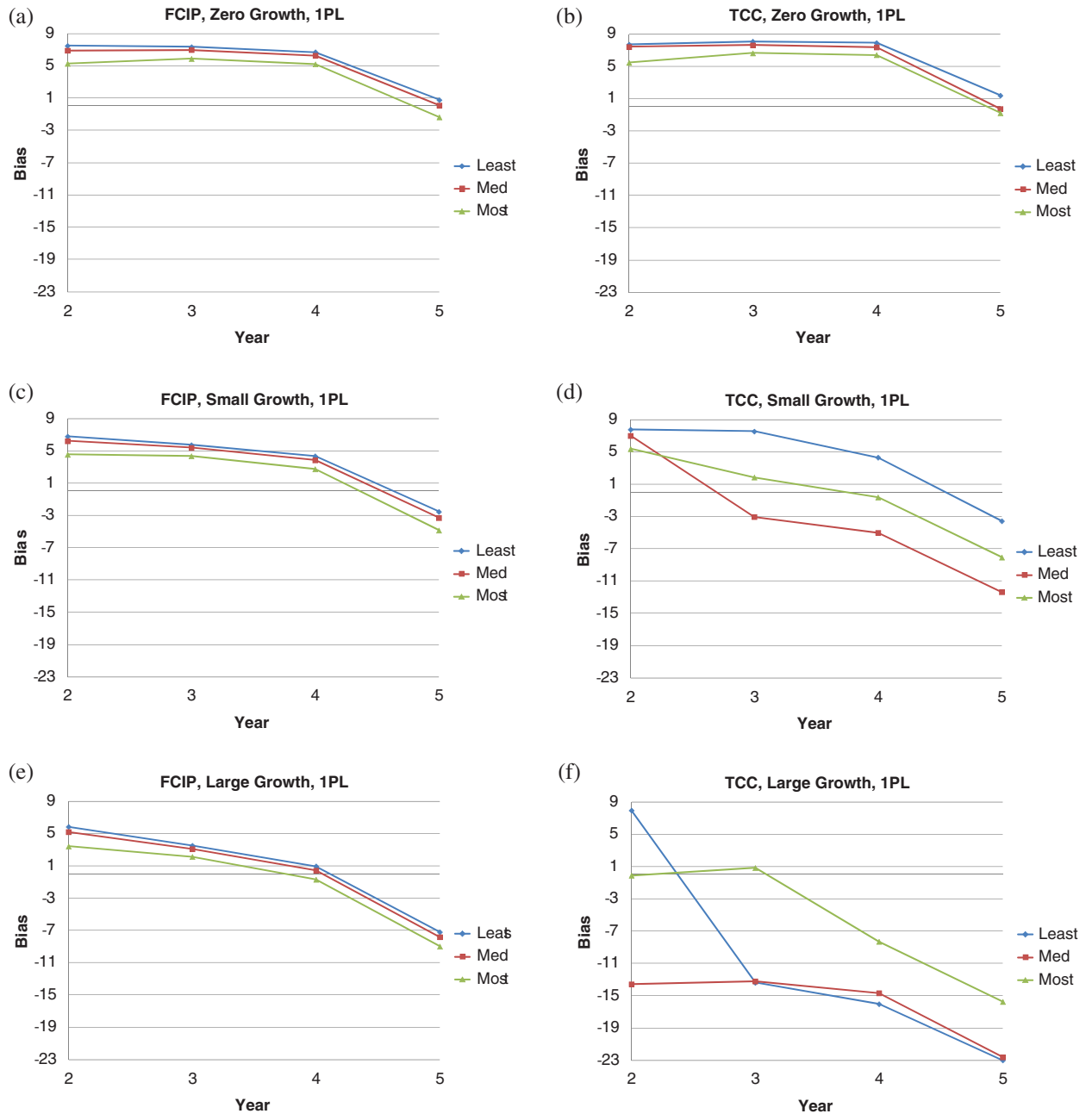

Figure 3 Mean bias by year under the one-parameter logistic model.

conditions. However, under 3PL and large growth, TCC had the largest average bias for the most stringent valid case criterion; the bias was negative, indicating that over time equating resulted in underestimation of student achievement.

Figures 5–8 compare the mean classification accuracy (over the 100 replications) in the equating years for different combinations of growth, valid case criterion, IRT model, and equating method. Results seemed to differ for the two IRT models. For 1PL, the most stringent valid case criterion was not necessarily associated with higher classification accuracy, and classification accuracy increased with time and with larger amounts of average growth. For 3PL, higher levels of stringency were generally associated with greater classification accuracy. Overall, the rates of classification accuracy were considerably higher for the most stringent valid case criterion than for the other two criteria, which had similar rates. However, the one exception was the large-growth condition with the TCC method, where classification accuracy dropped linearly from Year 3 forward, seeming to reflect the large increase in RMSE and bias under that

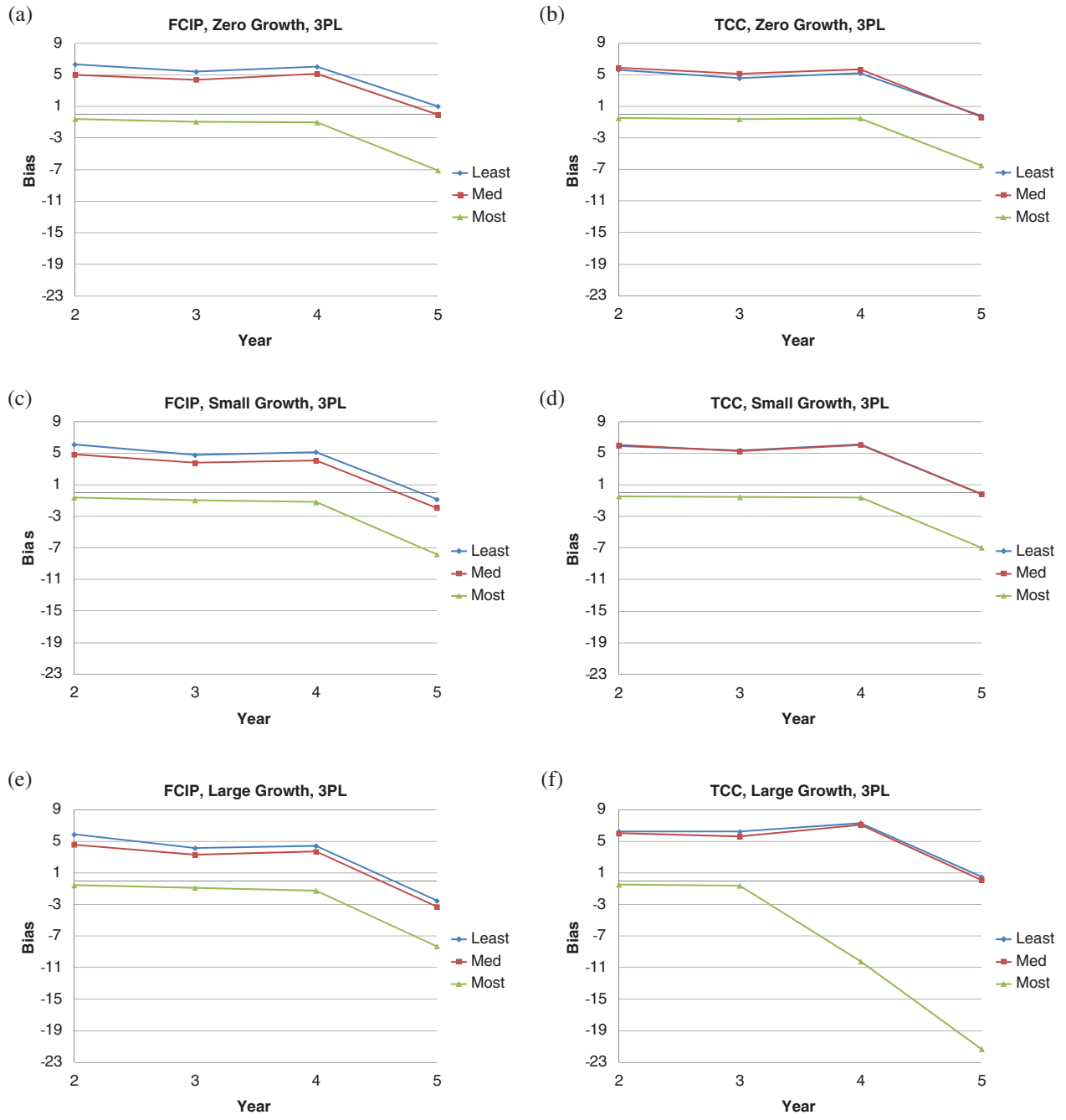

Figure 4 Mean bias by year under the three-parameter logistic model.

condition. If averaged across the 5 years, the classification accuracy rates were higher for the 3PL model than for the 1PL model.

Discussion

Equating is one of many important psychometric endeavors and is critical for maintaining comparability of test scores over time. When conducting equating, practitioners in the testing industry have often wondered and debated about the appropriate or optimal characteristics of the calibration/equating sample. What sample was used to derive the equating function is not always a matter of choice, as it is impacted by policy considerations as well as various operational constraints, such

Table 7 Average Mean Root Mean Square Error Across Years 2 – 5

Model	Equating method	Growth	Stringency of valid case criterion		
			Least	Med.	Most
1PL	FCIP	Zero	13.35	12.80	10.90
		Small	12.69	12.13	10.07
		Large	12.35	11.77	9.39
	TCC	Zero	7.12	7.17	6.70
		Small	8.37	11.89	9.45
		Large	18.61	18.99	13.30
3PL	FCIP	Zero	20.80	18.85	3.03
		Small	20.06	17.75	3.17
		Large	18.40	15.87	3.31
	TCC	Zero	16.49	17.84	2.94
		Small	17.29	17.83	2.90
		Large	17.43	16.63	8.78

Note. 1PL = one-parameter logistic model; 3PL = three-parameter logistic model; FCIP = fixed common item parameter method; TCC = test characteristic curve method.

Table 8 Average Mean Bias Across Years 2 – 5

Model	Equating method	Growth	Stringency of valid case criterion		
			Least	Med.	Most
1PL	FCIP	Zero	5.57	5.06	3.75
		Small	3.57	3.03	1.69
		Large	.77	.22	−1.03
	TCC	Zero	6.26	5.53	4.42
		Small	4.01	−3.35	−.37
		Large	−11.11	−16.01	−5.83
3PL	FCIP	Zero	4.69	3.61	−2.41
		Small	3.78	2.68	−2.66
		Large	2.97	2.07	−2.75
	TCC	Zero	3.79	4.08	−2.03
		Small	4.29	4.29	−2.15
		Large	5.06	4.69	−8.16

Note. 1PL = one-parameter logistic model; 3PL = three-parameter logistic model; FCIP = fixed common item parameter method; TCC = test characteristic curve method.

as test administration schedules and reporting timelines. One question involves whether the analysis sample must represent the testing population in terms of proportions of examinees with incomplete response strings, when policy requires that such responses be scored as incorrect, or given this policy context, will equating be more accurate if the calibration is based on students with complete response strings.

The purpose of this study was to examine how two commonly used IRT equating methods would perform over multiple administrations when calibration/equating samples varied as a function of different valid case inclusion stringencies. The general context is large-scale K – 12 assessment where policy often requires increasing levels of student achievement across years. The study extended the research on population invariance by examining the long-term effects of calibration sample valid case criterion on capturing growth under IRT equating.

Overall, study results indicate that the use of the most stringent valid case criterion would yield more accurate results under both zero-growth and growth conditions when omitted responses are coded as incorrect. The findings suggest that different levels of stringency in the valid case criterion will lead to meaningful differences (e.g., in terms of classification accuracy) in the equating results within one equating event as well as across multiple years of equating. These findings are

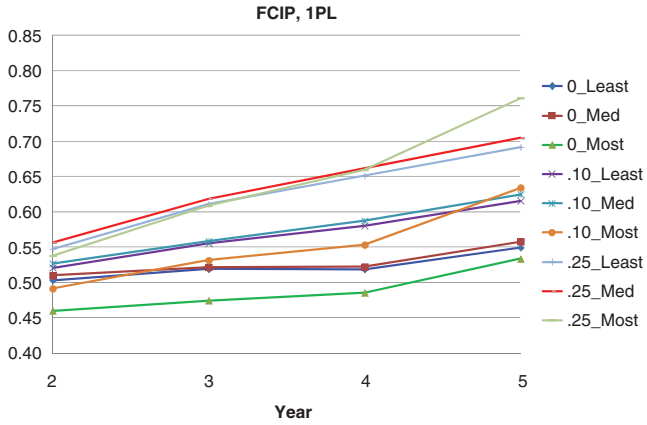

Figure 5 Mean classification accuracy for the fixed common item parameter method under the one-parameter logistic model.

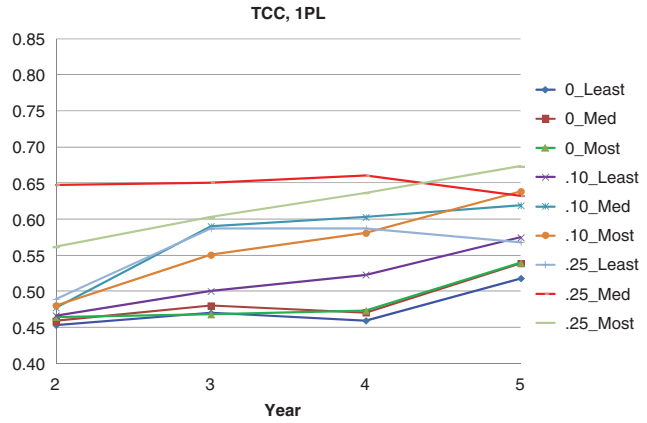

Figure 6 Mean classification accuracy for the test characteristic curve method under the one-parameter logistic model.

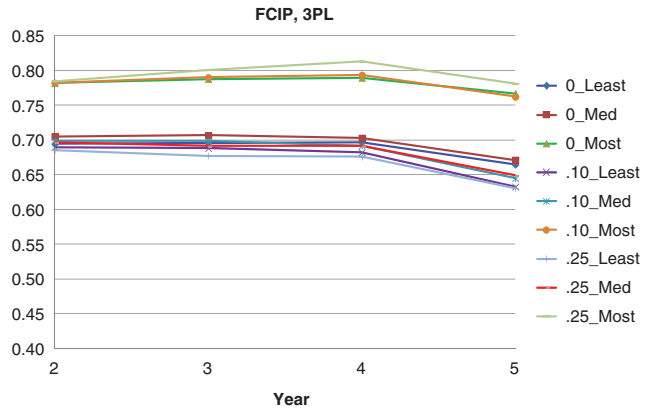

Figure 7 Mean classification accuracy for the fixed common item parameter method under the three-parameter logistic model.

consistent with Deng and Monfils (2011). Given that the test and anchor designs in this study were more reflective of operational realities in terms of item parameter values and test–anchor relationships than the tightly controlled within-form distribution of item difficulties of the previous study, the converging findings are very encouraging. These results are also consistent with those of Shin (2009), who found that coding missing responses as not presented rather than incorrect results in more accurate screening of items with parameter drift during equating, with results very similar to the baseline of complete data.

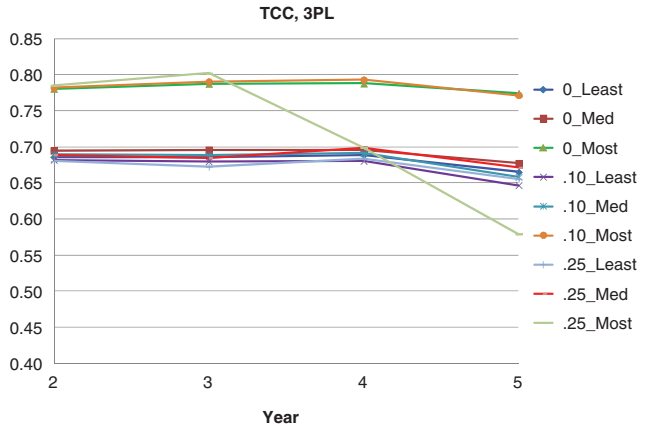

Figure 8 Mean classification accuracy for the test characteristic curve method under the three-parameter logistic model.

The reason why the most stringent valid case inclusion criterion led to generally the most accurate equating results over multiple administrations may be traced back to the assumptions required in applying IRT true score equating methods that were summarized by von Davier and Wilson (2007). One of the assumptions is that there are no omitted responses, which provides a theoretical basis for using the most stringent sampling criterion. Results of this study lend support to the importance of this assumption. With all test items completed, researchers have the most information available for calibration and equating, likely leading to the most accurate equating results.

TCC scaling did not work well for the large-growth condition regardless of the IRT model used: It had large RMSE and negative bias. This is not inconsistent with Keller, Skorupski, Swaminathan, and Jodoin (2004), who found that TCC underestimated growth when used with a long anchor, and Keller and Keller (2011), who concluded that although the characteristic curves methods, both TCC and ICC, would provide more robust theta estimates for a simple mean shift in the ability distribution, the FCIP would produce estimates with least bias for the more complex skew-shift condition, that is, when both mean and skewness in the ability distribution changed. Although we only examined mean shift here, the magnitude of the shift in the large-growth condition is .25, considerably larger than the .15 specified in Keller and Keller. As noted by Keller and Hambleton (2013), it is rare for effect sizes related to changes in the distribution to be larger than .10. From a study design point of view, we purposely used the .25 magnitude for large growth to test the boundaries of the long-term sustainability of IRT equating, knowing that it represented an extreme of population-level change in most educational settings. It is likely that the yearly shift of .25 posed a serious challenge for the TCC method.

The accumulation of bias over time in this study is also consistent with findings in the literature. Baldwin and Baldwin (2007) found that even in the zero-growth condition where the two samples were randomly equivalent, bias in the item parameter estimates increased as the administration number increased. Results of this study reflect a fairly consistent pattern of accumulated bias within model across years and conditions under FCIP, and that pattern was similar for 1PL and 3PL. However, as noted earlier, under TCC, the pattern of accumulated bias varied by model and conditions, with the greatest deviation under large growth.

Equating results may be less accurate when missing responses in the analysis data are treated as incorrect, due in part to increased error in estimating the item parameters in this context. For this study, the overall proportion of missingness was approximately 6%, which is not unusual in real operational testing data. Results of this study indicate that the different stringencies in valid case criteria can make a real difference in the ability of IRT equating to capture growth in the examinee distribution over multiple administrations. Given a policy context where omitted responses are scored as incorrect, the use of complete response strings for calibration and equating is more advantageous in that it produces more accurate equating results.

The accumulation of bias underscores the importance of monitoring the stability of the reporting scale over time, with periodic adjustments as needed done as part of scale maintenance. This is particularly important in contexts where population performance may shift due to factors such as changes in curriculum implementation or in stakes associated with test scores at the individual or group level.

Limitations and Future Research

This research had several limitations, and the results may not generalize beyond the conditions of the study design. As noted in the discussion, the generated test and anchor sets used in this study were designed to reflect operational realities in terms of item parameter values. Unlike Keller and Keller (2011), who used the same sets of items across all five administrations to eliminate test composition as a confounding source of error by design, there was some controlled difficulty variation in the forms and anchors sets used in this study. Future research should include form and anchor difficulty differences as a study variable.

Another limitation is that the simulated test had multiple-choice items only. Future research should look at mixed-format tests. Other factors that can be varied and examined include test and anchor length; sample size; and alignment, or lack thereof, of the test with examinee ability. Also, the growth in population ability may not be uniform across ability levels. Growth with a skewed distribution is worth investigating, especially with resources often prioritized to improve the performance of those at the lower end of the score distribution. In addition, one set of arbitrary cutscores was used in this study. To what extent the choice of cutscores may have impacted the results is unknown. Furthermore, the pattern and degree of missing responses may be varied as well.

Given the widespread concern in K–12 statewide testing with the growth of individual learners over school years, additional research is needed to study the impact of different levels of valid case inclusion stringency on measuring growth with vertical scaling.

As with all simulation studies, the findings of this study need to be evaluated with empirical data. Future research should compare how the stringencies in the valid case criterion affect equating in practical settings, where the composition of total test and anchor set reflect both content and statistical specifications.

Acknowledgments

The authors of this paper would like to thank Tammy Trierweiler for her contributions to this study. We are heartily appreciative of her help in setting up and running the code and her feedback and comments on a prior version of this paper.

Notes

1 To be concise, we will be using the term *growth* in this report to mean changes in student achievement at the population level, and not the growth at the individual level that is longitudinal in nature.

2 Although it is beyond the scope of this study, it should be noted that there is a body of research on IRT-based vertical scaling that has investigated the effect of IRT model choice on measuring growth at the individual level, dating back to concerns expressed in the 1980s over apparent scale shrinkage and more recently in the context of growth measures for educational accountability. As Briggs and Weeks (2009) have noted, IRT model choice is one of several design decisions that may impact the performance of vertical scales over time—these include, but are not limited to, IRT model, calibration and linking method, and ability estimation approach. Using empirical data, they found that, consistent with previous research, the magnitude of growth appeared to differ as a function of IRT model and linking method. In particular, they found that choice of IRT model has the greatest impact on scale variability (use of the 3PL/GPC models produced greater scale variability than use of the 1PL/PC models), but the choice of linking method and ability estimation has significant impacts as well.

3 It is a common design in K–12 testing to break a test into more than one section.

4 Consistent with most K–12 assessments, the multiple-choice items were right-scored, that is, scored as either correct or incorrect, without a penalty for guessing.

References

Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*, 213–234.

Baldwin, S., & Baldwin, P. (2007, April). *A comparison of IRT equating methods on recovering parameters and capturing growth in mixed-format te*sts. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Beguin, A. A. (2002). *Robustness of IRT test equating to violations of the representativeness of the common items in a nonequivalent groups design.* Unpublished manuscript.

Beguin, A. A. (2009, April). *The impact of violations of unidimensionality on IRT-linking: Follow up on the work on linking of Hanson and Béguin.* Unpublished PowerPoint presentation.

Brennan, R. L. (2008). A discussion of population invariance. *Applied Psychological Measurement*, *32*, 102–114.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*, 3–14.

DeMars, C. (2003, April). *Missing data and IRT item parameter estimation*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Retrieved from http://www.jmu.edu/assessment/wm_library/missdata.pdf

Deng, W., & Monfils, L. (2011, April). *Long-term effects of valid case inclusion criteria on capturing growth under IRT equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Psychological Measurement*, *3*, 37–52.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*, 225–245.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Han, K. T. (2012). *WinGen3: Windows software that generates IRT parameters and item responses* [Computer software]. Amherst, MA: University of Massachusetts–Amherst, Center for Educational Assessment. Retrieved from http://www.hantest.net/wingen

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*, 195–240.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *Journal of Experimental Education*, *71*, 229–250.

Keller, L. A., & Hambleton, R. K. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement*, *50*, 390–407.

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement*, *71*, 362–379.

Keller, L. A., Skorupski, W. P., Swaminathan, H., & Jodoin, M. G. (2004, April). *An evaluation of item response theory equating procedures for capturing changes in examinees distributions with mixed-format tests.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*, 25–41.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* [Computer software]. Iowa City: Iowa Testing Programs, University of Iowa.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, *45*, 225–245.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three Law School Admission Test administrations. *Applied Psychological Measurement*, *32*, 27–44.

Ludlow, L. H. & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59*, 615–630.

Muraki, E., & Bock, R. D. (1999). *Proprietary version of PARSCALE* [Computer software]. Chicago, IL: Scientific Software.

No Child Left Behind Act. (2001). 20 U.S.C. 6301 *et seq*.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). Sensitivity of equating results to different sampling strategies. *Applied Psychological Measurement*, *3*, 53–71.

Shin, S. (2009). How to treat omitted responses in Rasch model–based equating. *Practical Assessment, Research & Evaluation, 14*. Retrieved from http://pareonline.net/pdf/v14n1.pdf

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement*, *10*, 303–317.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true-score equating procedures* (Research Report No. RR-88-41). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210. https://doi.org/10.1002/j.2330-8516.1988.tb00297.x

von Davier, A. A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement*, *67*, 940–957.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement*, *32*, 11–26.

Yang, W. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement*, *41*, 33 – 42.

Yang, W., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination program examination. *Applied Psychological Measurement*, *32*, 45 – 61.

Yi, Q., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement*, *32*, 62 – 80.

Suggested citation:

Deng, W., & Monfils, L. (2017). *Long-term impact of valid case criteria on capturing population-level growth under item response theory equating* (Research Report No. RR-17-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12144

Action Editor: Gautam Puhan

Reviewers: Richard Schwarz and Zhen Wang