A Simple Parametric Procedure for Detecting Drift in Anchor Items

Xi Wang

Louis Roussos

Measured Progress

A Simple Parametric Procedure for Detecting Drift in Anchor Items

Abstract

A simple parametric statistic was recently proposed by Jiang, Roussos and Yu(2017) to detect drifting items based on the differences between the item response curves. Building on the previous study, this study proposed two methods for augmenting it with a standard error estimator. The enhanced statistical procedure is evaluated in a simulation study. Simulation results suggest the two standard error estimation methods produce very similar standard error estimates in both the type-I error and power conditions. These two methods also produce very similar standard error estimates with the empirical standard error.

Key words: item parameter drift, delta method, resampling method

Equating is a statistical process to adjust scores from test forms that differ in difficulty, so as to place scores on different tests on the same scale (Kolen & Brennan, 2014). One commonly used equating design is called non-equivalent-groups anchor test (NEAT) design, in which some items are re-used over time as anchors to link scores on different test administrations to the same scale. The accuracy of the linking relationship in the NEAT design depends on the extent to which the anchor items' parameters are invariant over time. Item parameter drift (IPD) could happen due to large position change of anchor items between test administrations (Meyers, Miller, & Way, 2009), and/or shift in school curriculum or instructional emphasis (Bock, Muraki, & Pfeiffenberger, 1988). As linking relationships could be negatively affected by IPD, some statistical procedures are typically conducted in practice to flag anchor items that have consequential IPD and exclude those items from the anchor set.

As IPD can be conceptualized as a special type of differential item functioning (DIF), DIF methods can be considered for IPD detection. However, in the linking context, not all

anchor items come from the same year's administration, making it difficult to use traditional nonparametric methods which combine item responses from different groups, such as Mantel-Haenszel (Holland & Thayer, 1988), SIBTEST (Shealy & Stout, 1993), and the standardization procedure (Dorans & Kulick, 1986). On the other hand, among parametric DIF or IPD methods that work with item parameter estimates directly, several other problems exist: (1) some procedures only focus on comparing individual parameters instead of item response functions (IRFs), such as Lord's χ^2 test (Lord, 1980), the robust z-approach (Huynh & Meyer, 2010); (2) some IRF-based procedures do not have analytically based critical values, such as the D^2 statistic in Wells, Hambleton, Kirkpatrick and Meng (2014); (3) some IRF-based statistics are hard to interpret, such as Raju's area measures (Raju, 1988; 1990).

Jiang, Roussos, and Yu (2017) proposed an IRF-based statistic that can be viewed as a parametric version of the estimators used in the standardization approach or SIBTEST, so it is easy to interpret and has known effect size. However, they did not derive a standard error associate with the statistic. Building on this recently developed statistic, we propose two methods to estimate its standard error in this study. A simulation study has also been conducted to evaluate the estimated standard error against the empirical standard error in both type-I error and power conditions. With the derived standard error, this statistic can be used as an inferential statistic in hypothesis testing, and it also overcomes the above-mentioned three limitations associated with existing IPD detection statistics.

A New Parametric IRF-based Statistic

Jiang et al. (2017) proposed an index, β, to summarize the difference in the IRFs between the reference (i.e., "year 1" – actual year can differ across items) and the focal group (year 2):

$$\beta_i = \int \frac{[E(Y_i|\theta, \boldsymbol{\delta_{iF}}) - E(Y_i|\theta, \boldsymbol{\delta_{iR}})]}{K_i} f(\theta) d\theta, \quad (1)$$

where Y_i is the score on item *i*, K_i is the maximum score, $\boldsymbol{\delta_{iF}}$ and $\boldsymbol{\delta_{iR}}$ are item parameter vectors in year 2 and 1, respectively. With dichotomous items, $E(Y_i|\theta, \boldsymbol{\delta_i}) = P(Y_i = 1|\theta, \boldsymbol{\delta_i})$; with polytomous items, $E(Y_i|\theta, \boldsymbol{\delta_i}) = \sum_{k=1}^{K_i} kP(Y_i = k|\theta, \boldsymbol{\delta_i})$; The division by maximum score allows β to be interpreted as a proportion-of-maximum score difference, and the integral averages this difference with respect to a common ability distribution. As β_i can be written as $\int \frac{E(Y_i|\theta, \boldsymbol{\delta_{iF}})}{K_i} f(\theta)d\theta - \int \frac{E(Y_i|\theta, \boldsymbol{\delta_{iR}})}{K_i} f(\theta)d\theta$, each term being the model-implied proportion correct for the item in a particular administration, β can be interpreted as the model-implied proportion correct difference in two administration given the same ability distribution.

To estimate β_i, a free (off-scale) calibration is conducted to estimate item parameters in year 2, i.e., $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}}$. Typically, year-1 operational-scale item parameters ($\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}$) are available from previous calibrations. Linking is conducted through the Stocking-Lord method (S-L; Stocking & Lord, 1983) to transform the off-scale year-2 parameters onto the operational year-1 scale. Linking constants are applied to $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}}$ to obtain $\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}$ on the same scale as $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}$**.** The choice of $f(\theta)$ can be based on any convenient distribution, such as the year-2 ability distribution. With regard to effect sizes, those of the standardization approach can be used: $|\beta| < 0.05$ implies negligible difference, $0.05 \leq |\beta| < 0.1$ implies moderate difference, and $|\beta| > 0.1$ implies large difference.

Standard Error Estimation Methods

Two methods are used to derive the standard error of the β estimator, i.e., $\hat{\beta}$. One method derives of an approximation of the standard error analytically through the multivariate delta method (Bickel & Doksum, 2015, p.319), and the other method constructs the standard error through a resampling approach. In the present study, only dichotomous items are considered, but the derivations can be easily generalized to polytomous items. The probability of a correct response is characterized by the three-parameter logistic model (3PLM):

$$P(Y_i = 1) = c_i + (1 - c_i)\frac{\exp(\alpha_i\theta + \beta_i)}{1 + \exp(\alpha_i\theta + \beta_i)} = c_i + (1 - c_i)\frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))} \quad (2)$$

where α_i and β_i are the slope and intercept parameters, a_i and b_i are the discrimination and difficulty parameters, and c_i is the pseudo-guessing parameter.

Delta Method

As $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}}$ and $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}$ are obtained from marginal maximum likelihood(MML), each follows a multivariate normal distribution asymptotically. Ignoring error in linking constants, $\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}$ is a linear transformation of $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}}$**,** and thus also follows an asymptotic multivariate normal distribution. As $\hat{\beta}$ can be written as the difference between two terms: a) $g\left(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}\right) = \int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}})f(\theta)d\theta$ and b) $g\left(\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}\right) = \int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}})f(\theta)d\theta$, each term being a function of a multivariate random variable, the delta method can be applied to calculate the asymptotic variance of each term. Specifically,

$$avar\left(g\left(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}\right)\right) = \frac{\partial g\left(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}\right)}{\partial \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}'} acov(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}) \frac{\partial g\left(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}\right)}{\partial \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}} \quad (3)$$

$$avar\left(g\left(\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}\right)\right) = \frac{\partial g\left(\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}\right)}{\partial \widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}'} acov(\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}) \frac{\partial g\left(\widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}\right)}{\partial \widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}}$$

The variance-covariance matrix for item parameter estimates can be estimated during item calibration. In this study, Flexmirt (Cai, 2012) is used as the item calibration software. As Flexmirt provides variance-covariance matrix for item slope (α_i), intercept (β_i) and logit of the guessing parameter (i.e., $g_i = log(\frac{c_i}{1-c_i})$), the following derivations are based on $\widehat{\boldsymbol{\delta}}_{\boldsymbol{i}} = (\alpha_i, \beta_i, g_i)'$**.**

As a free calibration is conducted for year-2, the asymptotic variance-covariance matrix estimated from the free calibration (i.e. $acov(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}})$) need to be transformed to get $acov(\widehat{\boldsymbol{\delta}}^{*}_{\boldsymbol{iF}})$. By S-L transformation, $\hat{a}^{*}_{iF} = \hat{a}_{iF}/A$, and $\hat{b}^{*}_{iF} = A\hat{b}_{iF} + B$, where A and B are linking constants.

Therefore, $(\hat{\hat{\alpha}}^{*}_{iF}, \hat{\hat{\beta}}^{*}_{iF}, \hat{\hat{g}}^{*}_{iF})' = T(\hat{\hat{\alpha}}_{iF}, \hat{\hat{\beta}}_{iF}, \hat{\hat{g}}_{iF})'$, where $T = \begin{pmatrix} \frac{1}{A} & 0 & 0 \\ -\frac{B}{A} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and thus

$acov(\hat{\hat{\boldsymbol{\delta}}}^{*}_{\mathbf{iF}}) = Tacov(\hat{\hat{\boldsymbol{\delta}}}_{\mathbf{iF}})T'$.

To calculate the partial derivative of $g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{i}})$ with respect to each parameter in $\widehat{\boldsymbol{\delta}}_{\boldsymbol{i}}$, by the Leibniz's rule for differentiation under the integral sign,

$$\frac{\partial g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{i}})}{\partial \widehat{\boldsymbol{\delta}}_{\boldsymbol{i}}} = \frac{\partial \int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}) f(\theta) d\theta}{\partial \widehat{\boldsymbol{\delta}}_{\boldsymbol{i}}} = \int \frac{\partial P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}})}{\partial \widehat{\boldsymbol{\delta}}_{\boldsymbol{i}}} f(\theta) d\theta. \quad (4)$$

Specifically,

$$\frac{\partial P(Y_i = 1)}{\partial \alpha_i} = (1 - L)L(1 - c_i)\theta$$

$$\frac{\partial P(Y_i = 1)}{\partial \beta_i} = (1 - L)L(1 - c_i) \quad (5)$$

$$\frac{\partial P(Y_i = 1)}{\partial g_i} = (1 - L)c_i(1 - c_i)$$

where $L = \frac{exp(\alpha_i\theta + \beta_i)}{1 + exp(\alpha_i\theta + \beta_i)}$. The integration can then be approximated by quadrature approximation.

Note that by treating the S-L linking constants as fixed values (i.e. ignoring the sampling error), $g(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}})$ and $g(\widehat{\boldsymbol{\delta}}^{*}_{\boldsymbol{iF}})$ are independent as the year-1 and year-2 calibration are usually conducted independently. Then the finite-sample variance of $\hat{\beta}_i$ is simply the same of the

variance of these two terms, i.e. var($\hat{\beta}_i$) $= var\left(g\left(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}\right)\right) + var\left(g\left(\widehat{\boldsymbol{\delta}}^{*}_{\boldsymbol{iF}}\right)\right) = \frac{avar\left(\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}\right)}{N_R} + \frac{avar\left(\widehat{\boldsymbol{\delta}}^{*}_{\boldsymbol{iF}}\right)}{N_F}$, where N_R and N_F are the calibration sample size for year-1 and year-2.

Resampling Method

As can be seen from Eq.1, $\hat{\beta}_i$ is a function of $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}$, $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}}$, and S-L linking constants *A* and *B* which are a function of the item parameter estimates for all anchor items in two years. Let $\widehat{\boldsymbol{\delta}}_R$ and $\widehat{\boldsymbol{\delta}}_F$ be the vector of item parameters for all items in year 1 and year 2 respectively, i.e., $\widehat{\boldsymbol{\delta}}_R = (\widehat{\boldsymbol{\delta}}_{1R}, \widehat{\boldsymbol{\delta}}_{2R}, \dots, \widehat{\boldsymbol{\delta}}_{IR})'$ and $\widehat{\boldsymbol{\delta}}_F = (\widehat{\boldsymbol{\delta}}_{1F}, \widehat{\boldsymbol{\delta}}_{2F}, \dots, \widehat{\boldsymbol{\delta}}_{IF})'$, where *I* is the total number of anchor items. The variability of $\hat{\beta}_i$ comes from the variability of $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}$, $\widehat{\boldsymbol{\delta}}_{\boldsymbol{iF}}$, $\widehat{\boldsymbol{\delta}}_R$ and $\widehat{\boldsymbol{\delta}}_F$. By the property of maximum likelihood estimators, the resampling method draws *J* samples $(\widehat{\boldsymbol{\delta}}^{(1)}_{\boldsymbol{iR}}, \dots, \widehat{\boldsymbol{\delta}}^{(J)}_{\boldsymbol{iR}}; \widehat{\boldsymbol{\delta}}^{(1)}_{\boldsymbol{iF}}, \dots, \widehat{\boldsymbol{\delta}}^{(J)}_{\boldsymbol{iF}}; \widehat{\boldsymbol{\delta}}^{(1)}_{\boldsymbol{R}}, \dots, \widehat{\boldsymbol{\delta}}^{(J)}_{\boldsymbol{R}}; \widehat{\boldsymbol{\delta}}^{(1)}_{\boldsymbol{F}}, \dots, \widehat{\boldsymbol{\delta}}^{(J)}_{\boldsymbol{F}})$ for each term from a multivariate normal distribution with the mean vector being the respective MML estimator of item parameters, and the dispersion matrix being the variance-covariance matrix for corresponding item parameters. The linking and calculation of $\hat{\beta}_i$ are then repeated for *J* times for each item, and the resulting $(\hat{\beta}_i^{(1)}, \dots, \hat{\beta}_i^{(J)})$ can be used to construct the empirical sampling distribution of $\hat{\beta}_i$ and to calculate the standard error of $\hat{\beta}_i$.

It can be seen that the resampling method overcomes one limitation in the delta method: In the delta method, the sampling error in S-L linking constants is ignored, whereas the resampling method takes that error into account. However, as the resampling method needs to repeatedly generate samples from multivariate normal distributions, it is more computationally intensive than the delta method. When there are a large number of anchor items, the distribution of $\widehat{\boldsymbol{\delta}}_R$ and $\widehat{\boldsymbol{\delta}}_F$ will be high dimensional, and thus the computation may be slow.

Simulation Design

A simulation study was conducted to compare the estimated standard error from the two methods with the empirical standard error in both type-I error and power conditions. In the Type-I error condition, a 50-item pre-equated test was simulated using the 3PLM to generate dichotomous item responses. True item parameters come from a large-scale state mathematics assessment. Two sample sizes, 3000 and 12000 were used, representing a small and a typical sample size respectively in a large-scale state educational assessment. The ability parameters for year-1 and year-2 samples was both simulated from $N(0,1)$.In the power condition, drift was simulated on the item difficulty parameter at two levels of prevalence: 20% and 40% of the items became easier in year-2. Only drift in item difficulty parameter was simulated as Jiang et al. (2017) found from a real-data analysis that the item discrimination parameter drift was small (< 0.2) in most items. Items with IPD were randomly sampled from the 50 items. IPD magnitude was simulated from two distributions, $U(0,0.4)$ (zero-to-small IPD) and $U(0,0.8)$ (zero-to-moderate). The $U(0,0.4)$ represented the drift amount in most of items from the real data analysis in Jiang et al. (2017), and $U(0,0.8)$ represented a worse situation. Unidirectional drift was simulated to mimic the worst impact of IPD on linking. Data generation was replicated 100 times.

For each replication, item calibration for each group was conducted in Flexmirt using MMLE. The variance-covariance matrix for item parameter estimates was estimated through the cross-product estimation. S-L constants were estimated in the R package "*plink*" (Weeks, 2010). With the resampling method, 100 samples were drawn from the multivariate normal distribution. In addition to estimating the standard error of $\hat{\beta}_i$ with the two methods in each replication, the empirical standard error of $\hat{\beta}_i$ was also obtained by calculating the standard deviation of $\hat{\beta}_i$ across replications.

Results

Table 1 presents summary information for $\hat{\beta}_i$ and the standard error estimate of $\hat{\beta}_i$ for items with and without drift in each condition. The last three columns in Table 1 summarize the empirical standard error of $\hat{\beta}_i$ averaged across all items, and the average standard error estimate using each method across all replications (i.e. estimated expected value of the standard error estimate) averaged across all items in each condition. The third column is a group indicator for items with and without drift: nonzero true β values indicate items with drift, and the fourth column shows the expected value of $\hat{\beta}$ averaged across items. First, as for the results on $\hat{\beta}$, the average expected value of $\hat{\beta}$ in the null condition suggest zero or very low bias in $\hat{\beta}$; however, in the power conditions, both the non-drifted and drifted items have negative bias in $\hat{\beta}$. Second, with respect to the results on standard error estimate of $\hat{\beta}$, the expected value of the standard error estimates are almost identical between the two methods in all conditions; compared to the empirical standard error, both methods result in slightly larger estimates, but the differences are not meaningful from a practical perspective.

Table 1. Average $\hat{\beta}$ and Standard Error Estimates for Items with and without Drift

N	Drift amount	True avg. β	avg.$\hat{\beta}$	avg. emp.SE	avg.SE (delta)	avg.SE (resampling)
3000	0	0	0.0000	0.0098	0.0115	0.0113
12000	0	0	0.0000	0.0050	0.0057	0.0061
3000	20%_U(0,0.4)*	0	-0.0065	0.0100	0.0117	0.0114
	20%_U(0,0.4)	0.0331	0.0277	0.0085	0.0099	0.0100
	20%_U(0,0.8)	0	-0.0119	0.0100	0.0119	0.0118
	20%_U(0,0.8)	0.0564	0.0478	0.0087	0.0100	0.0102
	40%_U(0,0.4)	0	-0.0120	0.0097	0.0113	0.0110
	40%_U(0,0.4)	0.0307	0.0181	0.0103	0.0121	0.0123
	40%_U(0,0.8)	0	-0.0269	0.0098	0.0114	0.0112
	40%_U(0,0.8)	0.0671	0.0410	0.0105	0.0120	0.0123
12000	20%_U(0,0.4)	0	-0.0069	0.0051	0.0058	0.0056
	20%_U(0,0.4)	0.0331	0.0276	0.0043	0.0049	0.0053
	20%_U(0,0.8)	0	-0.0118	0.0051	0.0059	0.0057
	20%_U(0,0.8)	0.0564	0.0476	0.0044	0.0049	0.0053

40%_U(0,0.4)	0	-0.0122	0.0048	0.0055	0.0054
40%_U(0,0.4)	0.0307	0.0178	0.0051	0.0059	0.0058
40%_U(0,0.8)	0	-0.0267	0.0050	0.0056	0.0056
40%_U(0,0.8)	0.0671	0.0405	0.0053	0.0059	0.0059

* 20%_U(0,0.4) represents the condition where there are 20% drifted items, and the IPD amount is sampled from Uniform(0,0.4).

Figures 1 through 5 show the results by individual items in each condition. In each figure, the expected value of $\hat{\beta}$ is represented by the black dots, the broken bars above dots represent the expected value of the standard error estimate of $\hat{\beta}$ by each method, and the solid bars below the dots represent the empirical standard error. The red horizontal line represents the true β for each item. Items on each plot are ordered in groups by their discrimination parameters first, and then within each group (every 10 items), items are ordered by their difficulty. Figure 1 displays the results in the null conditions, and Figures 2 to 5 display the results in the power conditions.

As Figure 1 shows, the expected value of $\hat{\beta}$ is very close to the true β in the null condition for all items, and the expected value of the standard error estimate is very close to the empirical standard error in both methods. The standard error slightly increases with the item difficulty, but does not seem to have a clear pattern with item discrimination. The standard error estimate under the sample size of 3,000 is almost twice than that under the sample size of 12,000. This is as expected, because standard error is inversely related to the square root of the sample size. Under the power conditions, there is a negative bias in $\hat{\beta}$ in both drifted and non-drifted items, as a result of simulated negative change in year-2 item difficulty. The bias is negative for non-drifted items because the S-L transformation makes the year-2 post-equated parameters appear more difficult than the corresponding year-1 values, and thus $\int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}) f(\theta) d\theta < \int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}) f(\theta) d\theta$. In contrast, for the drifted items, the S-L transformation makes the year-2 item difficulty more similar to their corresponding year-1 values, so the difference between $\int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}^*_{\boldsymbol{iF}}) f(\theta) d\theta$ and $\int P(Y_i = 1|\theta, \widehat{\boldsymbol{\delta}}_{\boldsymbol{iR}}) f(\theta) d\theta$ is smaller

than the difference between $\int P(Y_i = 1|\theta, \boldsymbol{\delta_{iF}})f(\theta)d\theta$ and $\int P(Y_i = 1|\theta, \boldsymbol{\delta_{iR}})f(\theta)d\theta$. As for the standard error estimate in the power conditions, the expected value of the standard error estimate is very similar to the empirical standard error in both methods in all power conditions.

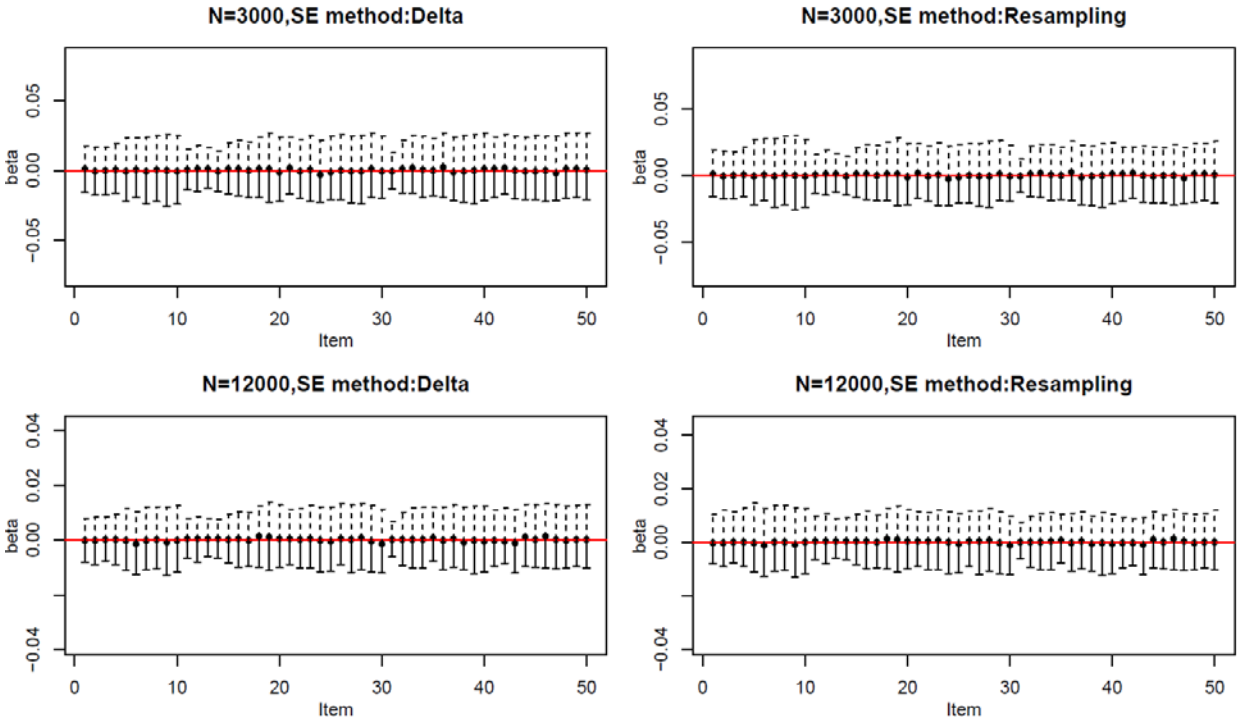

Figure 1. Expected value of $\hat{\beta}$ and standard error estimates in the null conditions.

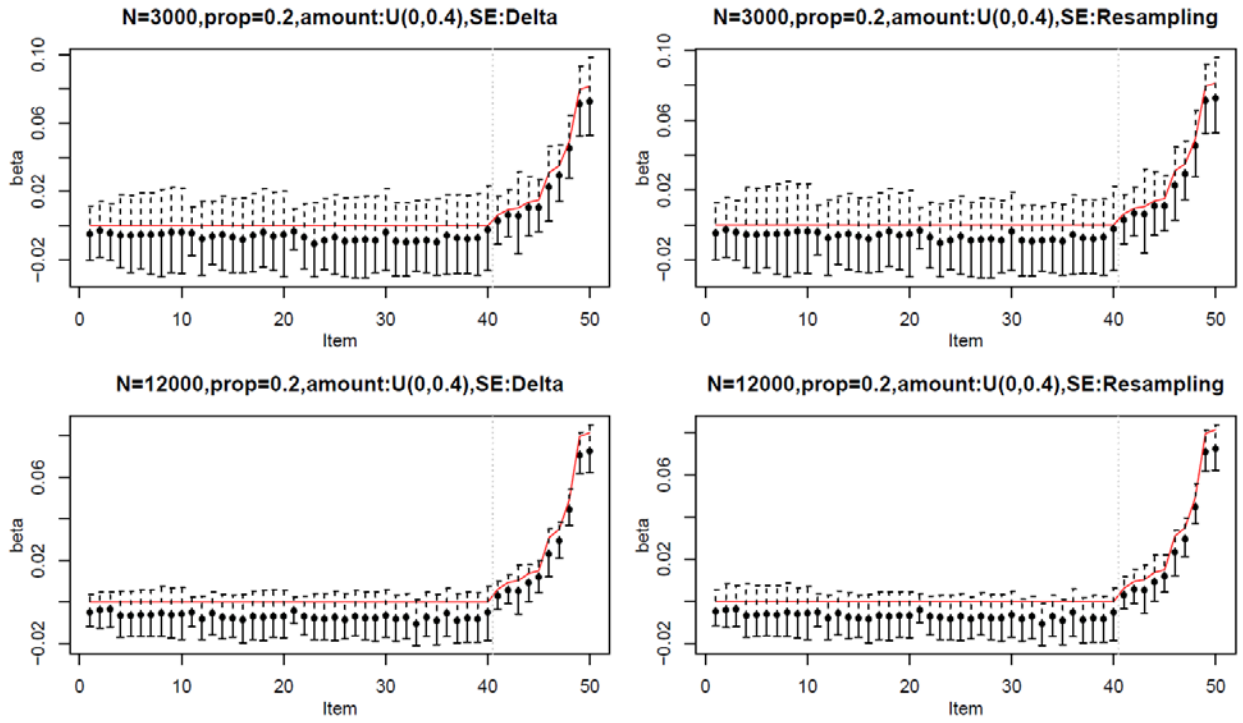

Figure 2. Expected value of $\hat{\beta}$ and standard error estimates in the power condition with 20% drifted items, and drift amount sampled from *Uniform*(0,0.4). The vertical dotted lines separate items without (Left) and with drift (Right).

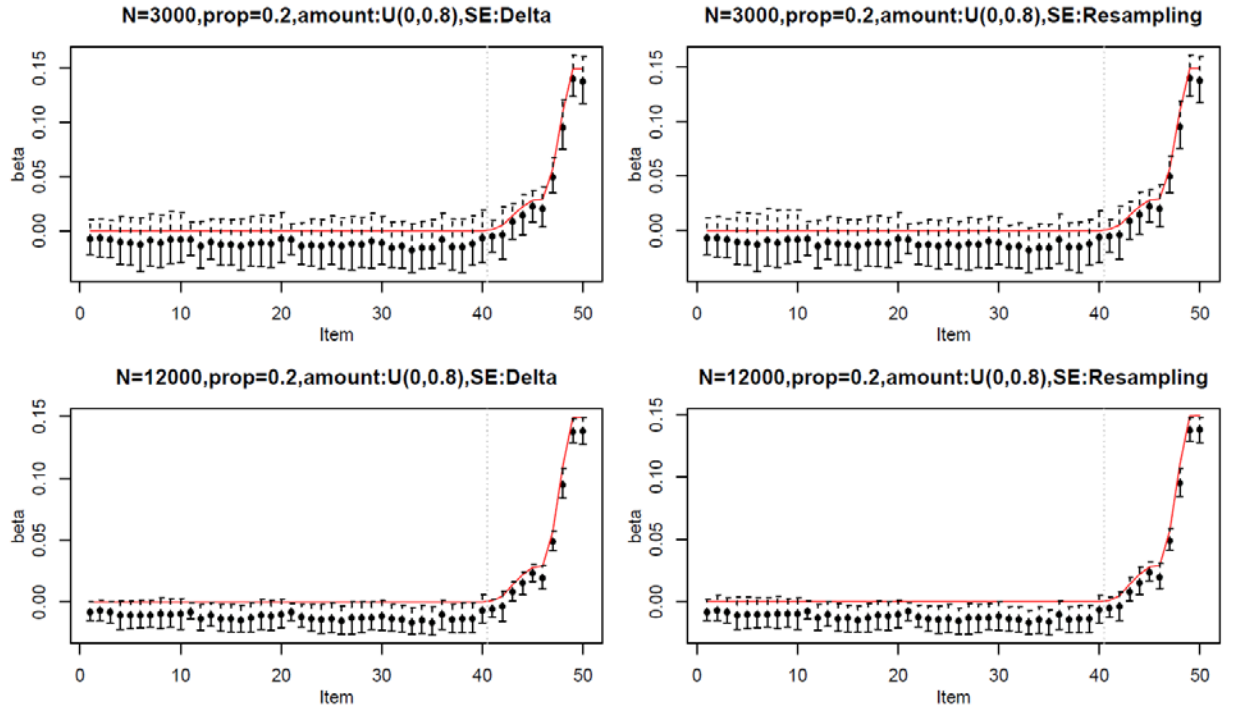

Figure 3. Expected value of $\hat{\beta}$ and standard error estimates in the power condition with 20% drifted items, and drift amount sampled from *Uniform*(0,0.8).

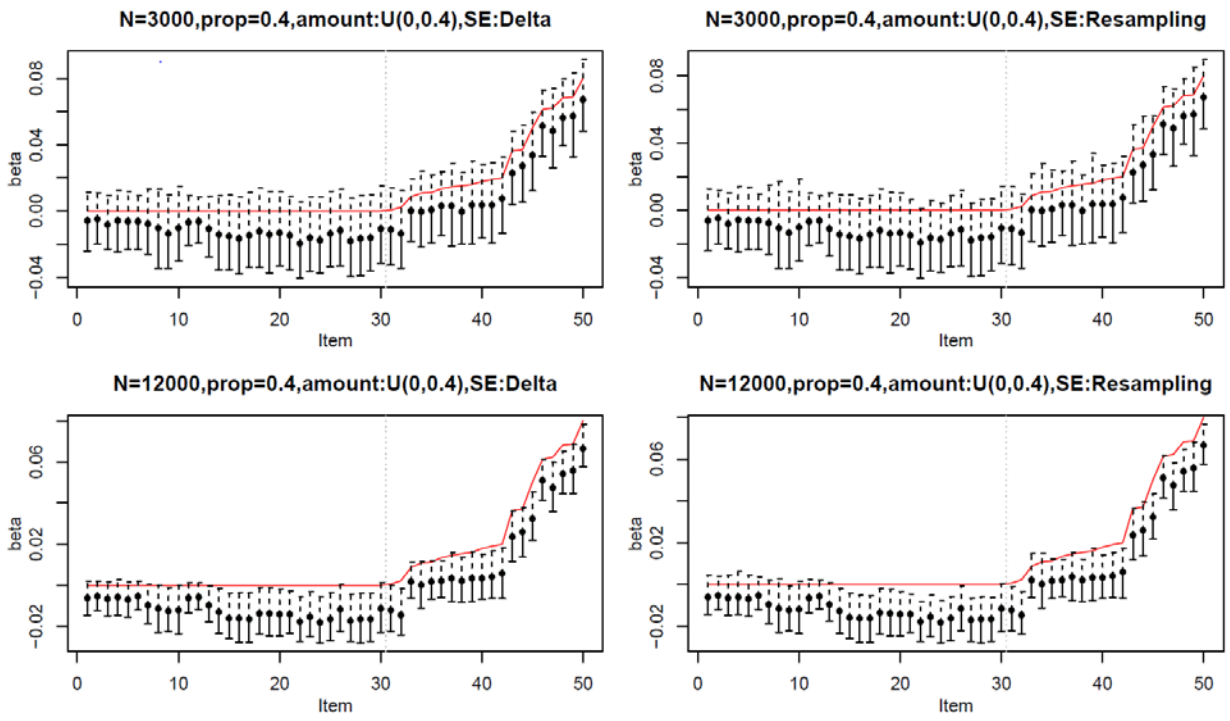

Figure 4. Expected value of $\hat{\beta}$ and standard error estimates in the power condition with 40% drifted items, and drift amount sampled from *Uniform*(0,0.4).

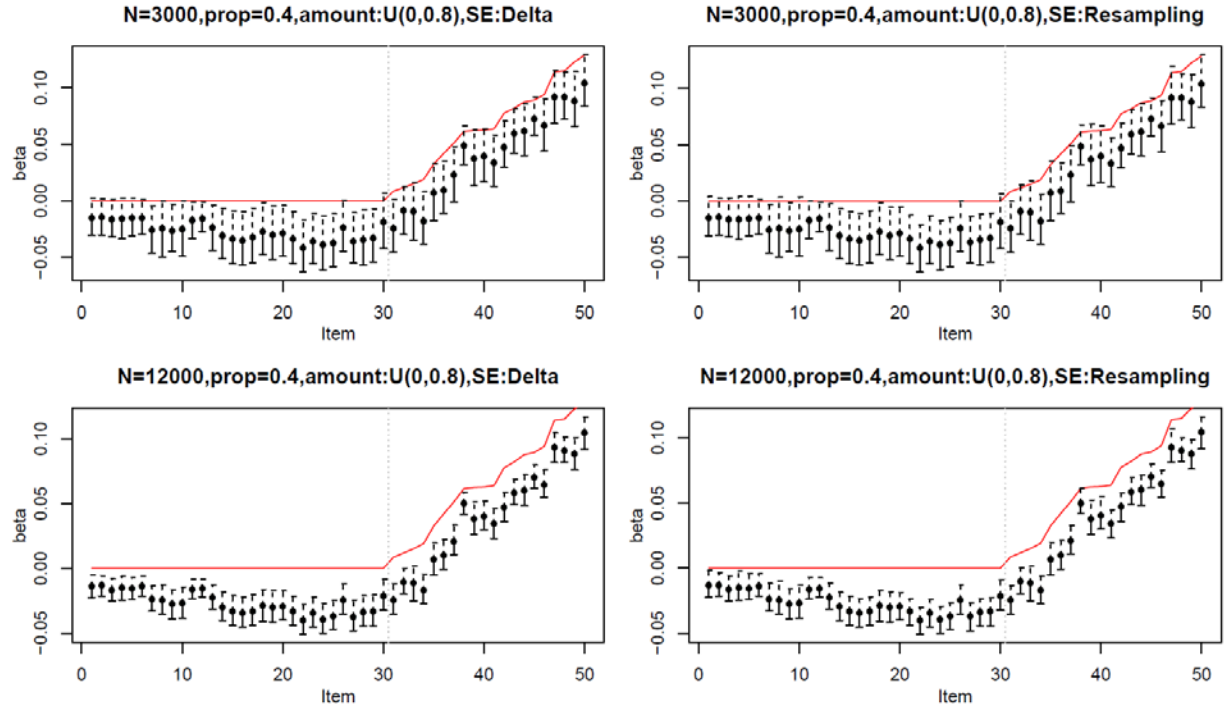

Figure 5. Expected value of $\hat{\beta}$ and standard error estimates in the power condition with 40% drifted items, and drift amount sampled from *Uniform*(0,0.8).

Discussions

As the results suggest, the delta method and resampling method can produce very similar standard error estimates for $\hat{\beta}$. Even though the delta method ignores the error in linking constants, its practical impact is very small. Although both methods produce slightly larger standard error estimates than the empirical standard error, the difference is not consequential in a practical sense. These results provide empirical support for using the two methods to calculate the standard error in practice. With the standard error being calculated, the $\hat{\beta}$ can be used as an inferential statistic to conduct hypothesis testing.

Another important implication of the simulation results is that the standard error of $\hat{\beta}$ is no larger than 0.025 among most items in all simulated conditions, even with sample size of 3,000. This justifies the choice of effect size for flagging and interpreting β. As 0.05 is used as the threshold for negligible difference, if $\hat{\beta} > 0.05$ is observed, we can have more confidence in concluding that this difference is caused by a systematic change in item parameters rather than simply due to sampling error.

Lastly, as the results implied, the $\hat{\beta}$ for non-drifted items could deviate from 0 significantly when a lot of items have large drift, due to the impact of IPD on linking constants. So some procedures should be conducted to remove drifted items from the anchor set. The current simulation results suggest that the non-drifted items tend to have similar $\hat{\beta}$ values, and the $\hat{\beta}$ values for drifted items tend to have a different pattern from that of the non-drifted items. Therefore, a clustering technique could be applied to separate these two groups of items, so that only non-drifted items are used in linking. This could be a future research direction.

References

Bickel, P.J., & Doksum K.A. (2015). *Mathematical statistics: basic ideas and selected topics, Volume I* (2nd ed.). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Bock, R. D., Murakl, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Holland, P. W., & Thayers, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure (Technical Rep. No. 86-69). Princeton, NJ: Educational Testing Service.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment Research & Evaluation, 15*(2), 1-8.

Jiang, J., Roussos, L., & Yu, L. (2017, April). *An iterative procedure to detect item parameter drift in equating items*. Paper presented at the National Council on Measurement in Education Conference, San Antonio, TX.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas betweentwo item response functions. *Applied Psychological Measurement, 14,* 197-207.

Shealy, R., & Stout, W. (1993). A model-based approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Weeks, J.P. (2010). plink: an R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1-33.

Wells, C.S., Hambleton, R.K., Kirkpatric, R. & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27*(3), 214-231.