CollegeBoard
inspiring minds™

Research Report
No. 2010-2

An Investigation of Scale Drift for Arithmetic Assessment of ACCUPLACER®

Hui Deng and Gerald Melican

www.collegeboard.com

College Board Research Report No. 2010-2

An Investigation of Scale Drift for Arithmetic Assessment of ACCUPLACER®

Hui Deng and Gerald Melican

The College Board, New York, 2010

Hui Deng is a senior psychometrician at the College Board.

Gerald Melican is chief psychometrician at the College Board.

The College Board

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the College Board is composed of more than 5,700 schools, colleges, universities and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,800 colleges through major programs and services in college readiness, college admission, guidance, assessment, financial aid and enrollment. Among its widely recognized programs are the SAT®, the PSAT/NMSQT®, the Advanced Placement Program® (AP®), SpringBoard® and ACCUPLACER®. The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities and concerns.

For further information, visit www.collegeboard.com.

Printed in the United States of America.

Contents

Tables

Figures

Objectives of Inquiry and Theoretical Background

For IRT-based testing programs that administer multiple test forms over time, it is critical to maintain a stable reporting scale so that scores are comparable across years, administrations and test forms. According to the invariance property of IRT (Baker, 1992; Hambleton & Swaminathan, 1985), item parameters estimated from different samples of the same population are invariant; this feature has been widely used for test equating, score scaling and online calibration in the CAT environment. However, changes in item parameters are likely to occur due to changes in curriculum, frequent exposure of items or other reasons. Such changes can threaten the validity of test scores by introducing trait-irrelevant differences on ability estimates. For example, to the extent that an item is known due to overexposure, the item becomes easier and less discriminating, causing errors in proficiency estimation using the original item parameters. Changes in parameter values for different subgroups have been referred to as differential item functioning (DIF) (Holland & Wainer, 1993), while changes across testing time have been referred to as item parameter drift (IPD) (Goldstein, 1983; Bock, Muraki & Pfeiffenberger, 1988). Wise and Kingsbury (2000) indicate that it is appropriate to perform scale drift studies by recalibrating some previously calibrated items to ensure the drift is nondirectional and within bounds that are expected due to sampling error.

IPD has been extensively studied for paper-and-pencil tests. Most studies using two time points found IPD had minor impact on the resulting ability estimates (Wells, Subkoviak & Serlin, 2002; Rupp & Zumbo, 2003a, 2003b). However, there has been concern that item drift may compound over time, especially if drifting items are used in linking of test forms (Kim & Cohen, 1992). If an item bank is not monitored for drift over years, it is likely that the percentage of drifting items as well as the magnitude of the drift may accumulate over time and have detrimental effects on the measurement of the intended construct. A few studies have examined item parameter drift over multiple test occasions for paper-and-pencil tests (Chan, Drasgow & Sawin, 1999; DeMars, 2004). The study by Wollack, Sung and Kang (2006) suggests that the choice of linking model can have large impact on the effects of IPD on theta estimates and passing rates.

Despite the fact that successful implementation of CAT depends on the integrity of its item pool and stability of the item parameters, the issue of IPD has been scarcely investigated in the context of CAT, and very few relevant studies could be found in the literature. In a simulation study, Stocking (1988) found evidence of scale drift through various rounds of simulations for online-calibration. Guo and Wang (2005) compared item parameters for pretest items calibrated at two time points using test characteristics to measure the scale drift at the test level. However, no study has systematically examined scale drift at the item level across multiple years in the context of CAT.

The current study was designed to extend the current literature to study scale drift in CAT as part of improving quality control and calibration process for ACCUPLACER®, a battery of large-scale adaptive placement tests. The study aims to evaluate item parameter drift using empirical data that span four years from the ACCUPLACER Arithmetic assessment.

Method

Data Source

The data and item calibration procedure used in the study were based on ACCUPLACER, which consists of adaptive tests designed to measure reading, writing and mathematics skills. The tests are delivered online, and the test lengths range from 12 to 20 multiple-choice items. The scores from the tests are used for course placement and assessment of academic progress. The test chosen for the current study is the ACCUPLACER Arithmetic test, which measures the ability to perform basic arithmetic operations and to solve problems that involve fundamental arithmetic concepts. The 17 questions on the Arithmetic test are divided into three types: operations with whole numbers and fractions, operations with decimals and percents, and applications and problem solving. The test is untimed, administered adaptively under 32 constraints with respect to content category, item property and key distribution.

The Arithmetic item pool was refreshed in January 2004 by calibrating and scaling pretest items and adding to the item pool. The refreshed item pool has item parameters on the reference scale and has been used for CAT administrations in the subsequent years through 2007. The data used in the current study are based on the operational data from years 2004, 2005, 2006 and 2007 administrations. After data cleaning, the yearly datasets contain 805,943; 871,223; 947,727 and 993,575 records for 2004, 2005, 2006 and 2007 samples, respectively. Summary statistics of item exposure rates for the 223 items in the yearly samples are included in the Appendix. The maximum exposure rate in each yearly sample is as high as 0.55, meaning 55 percent of test-takers in that year have seen the item. As the tests from 2004

through 2007 were based on the old CAT system for the assessment, which did not implement exposure control, there has been concern for whether certain overexposed items would exhibit parameter drift. This is a major reason for this study.

Each yearly dataset was calibrated separately to investigate item parameter drift. Because item parameter estimates from two separate calibrations could involve calibration and scaling error, multiple calibrations were replicated first with data from 2004 in order to estimate the magnitude of random variation in the calibration process and also in order to establish empirical criterion for evaluating item parameters estimates based on the 2005, 2006 and 2007 calibrations.

Calibration

As CAT by design can administer different items for different examinees by targeting at the individuals' ability estimates, the dataset for each year's calibration was a sparse matrix, with each examinee providing answers to a set of 17 items. The remaining items the examinee did not take were treated as not-presented during calibration. Only items to which 200 or more examinees responded in each yearly dataset were included for calibration, which resulted in 223 items being calibrated and examined for item parameter drift.

The calibrations were based on 3PL model and carried out using Bilog-MG, which implements marginal maximum a posteriori estimation procedure for estimating item parameters (Bock & Aitkin, 1981). Because of the nature of the data, it was necessary to set priors when calibrating item parameters to obtain a converged solution. For each item, the baseline *a* and *b* parameters for each item were used to set prior means for estimating *a* and *b* parameters, while the default prior standard deviations of 0.50 for the log of *a*'s and 2.00 for the *b*'s were used. For c-parameter, the default priors from Bilog were used for all items.

The item parameters calibrated for each year were transformed to the baseline scale using mean/sigma transformation. The formulas used for obtaining the slope (A) and intercept (B) for the transformation are as follows (Kolen & Brennan, 1995):

$$A = \frac{\sigma(b_a)}{\sigma(b_c)} \tag{1}$$

$$B = \mu(b_a) - A\mu(b_c) \tag{2}$$

where *a* represents the baseline parameters and *c* represents parameters resulted from each yearly calibration. μ's and σ's were based on items with valid parameter estimates in the yearly calibration. There were several items that were not calibrated due to negative item-total correlation and Bilog produced parameter estimates for these items based on their priors. Because these parameter estimates were far from being accurate, they were excluded from computing transformation constants.

The item parameter estimates from each yearly calibration were then transformed using the following equations (Kolen & Brennan, 1995):

$$a_{aj} = \frac{a_{cj}}{A} \tag{3}$$

$$b_{aj} = Ab_{cj} + B \tag{4}$$

$$C_{aj} = C_{aj} \tag{5}$$

where a_{aj}, b_{aj} and c_{aj} are the item parameter estimates for item *j* on baseline scale, and a_{cj}, b_{cj} and c_{cj} are the item parameter estimates for item *j* on the scale from each yearly calibration.

Empirical Evaluation of Item Parameter Drift

The problem of identifying IPD is statistically identical to that of identifying DIF. While DIF analyses attempt to examine whether an item functions differentially across examinee subgroups, and IPD analyses attempt to examine whether an item functions differentially across testing time, the underlying question is the same. The current study used the nonconfirmatory differential item functioning (NCDIF) index developed by Raju, van der Linden and Fleer in 1995 to examine item parameter drift at the item level:

$$NCDIF_i = E_F\left(P_{iF}(\Theta) - P_{iR}(\Theta)\right)^2 = E_F d_i^2 \tag{6}$$

where P_{iF} and P_{iR} are the probability of a correct response at a given theta level using item parameter estimates from the reference group and the focal group, respectively, and d_i refers to the difference in probability for item *i*. The NCDIF is based on the assumption that all items in the test are free of DIF except for the item being studied, which corresponds to most of the IRT-based DIF methods.

In order to set up a null distribution of the NCDIF values for each item, random samples of 15,000 records were drawn without replacement from the 2004 operational data, using different random seeds each time to generate different response matrices. Each sample was then independently calibrated to estimate item parameters for the 223 items under study, and scaled to the baseline scale using mean/sigma transformation. Fifty-three replications were done for the current study. To evaluate the amount of estimation errors at item level,

the transformed item parameter estimates based on samples 2 through 53 were compared to that based on sample 1, and the NCDIF was computed for comparison of calibration results from each sample against sample 1. In the context of this study, sample 1 was treated as a reference group, and samples 2 through 53 were each treated as a focal group.

For each item, with the NCDIF values computed using item parameter estimates from the replication samples, the 90th percentile was obtained from the empirical distribution of the NCDIF and used as a cut-off to evaluate item parameter drift in the real data calibration based on the yearly sample. Because the sample size for the replication study — 15,000 — is smaller than the yearly operational sample size, the NCDIF criterion may be on the conservative side. Consequently, the type I error rate of 0.10 instead of 0.05 was chosen.

For item parameters calibrated using 2004, 2005, 2006 and 2007 data, the item characteristic curves and test characteristic curves were plotted and examined using parameters resulted from each yearly calibration. To identify drifted items, the NCDIF was computed for each item using 2004 sample as a reference group, and each subsequent year's sample as a focal group. For each item, the NCDIF values for each of the 2005/ 2004, 2006/ 2004 and 2007/ 2004 comparisons were compared to the empirical cut-score for the corresponding item, and items with NCDIF values larger than the cut-score were flagged as drifted.

Results

For the yearly calibrations, several items could not be calibrated due to the negative item-total correlation in the yearly sample. Item 219 had a negative biserial in each yearly sample of 04, 05, 06 and 07 and was not calibrated. In addition, item 18 had a negative biserial in the 2006 sample and was not calibrated for the 2006 sample.

Table 1 presents mean and standard deviation of item parameter estimates based on each yearly calibration. Given that the mean and SD of *b* parameter estimates were equated in the mean/sigma transformation, the means and SDs of *b* estimates were essentially equal across years, as expected. Means and SDs for the *c* parameters are also highly consistent across years, with the 2006 and 2007 estimates having slightly higher means (by 0.01) than the 2004 and 2005 estimates. The *a* parameter estimates were less consistent across years. The means range from 1.70 to 2.25, and the SDs range from 2.71 to 3.67. The mean and SD of *a* parameters from 2005 were substantially higher compared to those of the other years, while the mean and SD from the 2007 calibration were lower than the other years. The *a* estimates from the 2004 and 2006 calibrations were relatively more consistent.

Table 1

Descriptive Statistics of Item Parameter Estimates Across Years

Year	*N Items*	*a*		*b*		*c*	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
2004	222	1.84	3.00	-0.34	1.35	0.09	0.09
2005	222	2.25	3.67	-0.34	1.35	0.09	0.09
2006	221	1.81	2.91	-0.33	1.35	0.10	0.09
2007	222	1.70	2.71	-0.34	1.35	0.10	0.09

The correlations among item parameter estimates calibrated using the yearly samples are shown in Table 2. The bi-year correlations for *a*, *b*, *c* parameter estimates were all consistently high. The correlations range from .989 to .998 for *a* parameters, from .983 to .998 for *b* parameters and from .953 to .977 for *c* parameters.

Table 2

Correlations of Item Parameter Estimates Across Years

Parameter	*Year*	*2004*	*2005*	*2006*	*2007*
a	2004	1.000	0.993	0.989	0.991
	2005	0.993	1.000	0.995	0.998
	2006	0.989	0.995	1.000	0.996
	2007	0.991	0.998	0.996	1.000
b	2004	1.000	0.995	0.997	0.994
	2005	0.995	1.000	0.990	0.983
	2006	0.997	0.990	1.000	0.998
	2007	0.994	0.983	0.998	1.000
c	2004	1.000	0.971	0.953	0.953
	2005	0.971	1.000	0.976	0.977
	2006	0.953	0.976	1.000	0.972
	2007	0.953	0.977	0.972	1.000

Figure 1 shows test characteristic curves (TCC) across the yearly calibrations from 2004 through 2007, represented by total proportion correct on the set of items studied. Overall, the TCCs are parallel and reasonably similar across years. However, the 2005 curve seems to be more deviant than the curves from the other years. At the ability range of -0.3 to -3, the test seems to be more difficult in 2005. This is because a lower proportion of total items correct is associated with the same ability in that range compared to the other years. On the contrary,

the test seems to become easier in 2005 for the ability range of 0 to 2. This is because students with the same ability tended to receive higher scores in 2005 compared to the other years. However, the biggest difference in TCC, in either direction, is only about 0.03 in the proportion correct metric.

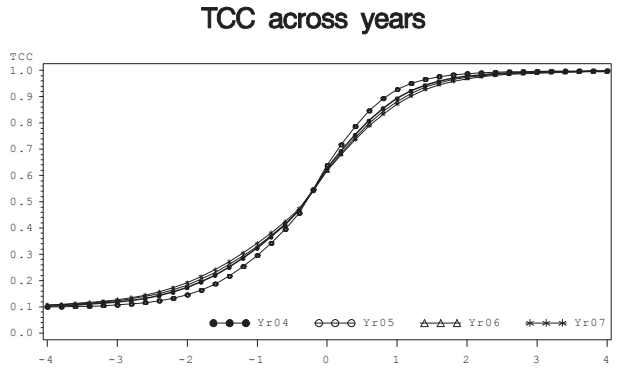

Figure 1. Test characteristic curves across years.

The individual item characteristic curves (ICCs) have been depicted for each item based on yearly calibrations, with four curves presented on the same graph to allow direct comparison of ICCs obtained from different years. Figure 2 shows an item whose ICCs are highly similar as estimated from 2004 through 2007. The majority of items being studied had ICCs resembling this type of similarity across the yearly calibrations.

Out of the 222 items each with NCDIF values from 2005, 2006 and 2007 calibrations compared to the item's NCDIF criterion, only two items were flagged as showing parameter drift. The ICCs for these two items are shown in Figure 3 and Figure 4. Table 3 presents the item parameter estimates and NCDIF results based on yearly calibrations for these two items, as well as the NCDIF criterion values computed from the within-2004 replications.

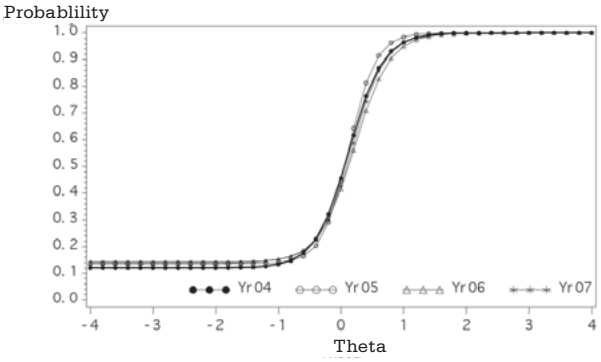

Figure 2. An item with highly consistent ICCs across years.

Based on the ICCs and item parameter estimates across years, item M_070445, ICCs from 2005, 2006 and 2006 exhibit large differences compared with the ICC from the base year 2004. This item became harder and more discriminating after 2004, and it was flagged as drifted in all years after 2004.

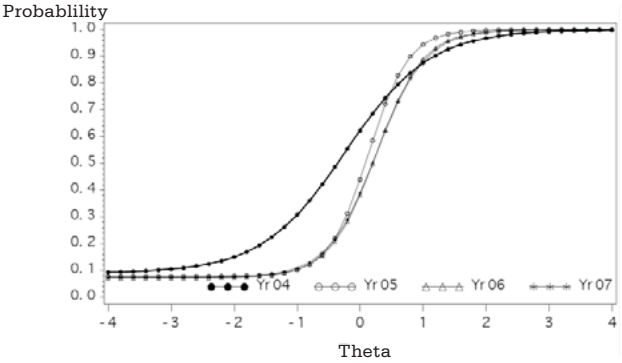

Figure 3. ICCs for item M_070445, which was flagged for IPD.

Table 3

Item Parameters and NCDIF Results for Flagged Items

Item Num	*Item ID*	*Year*	*a*	*b*	*c*	*NCDIF*	*NCDIF Criterion*
ITEM115	M_070445	2004	0.88	-0.23	0.09		0.0085
		2005	1.87	0.14	0.08	0.0240	
		2006	1.59	0.26	0.07	0.0289	
		2007	1.52	0.26	0.07	0.0273	
ITEM144	R_2ARIT0547B	2004	1.28	0.12	0.12		0.0084
		2005	1.86	0.14	0.20	0.0027	
		2006	1.84	0.42	0.30	0.0121	
		2007	1.69	0.41	0.31	0.0135	

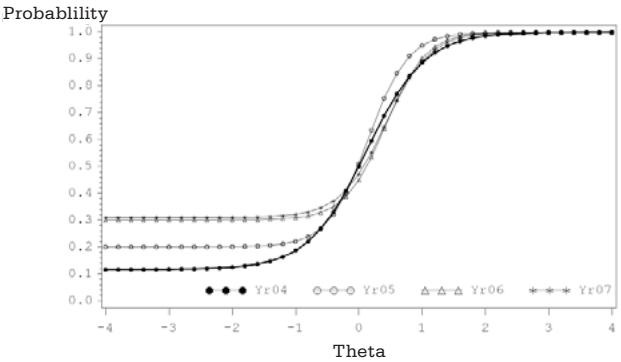

Figure 4. ICCs for item R_2ARIT0547B, which was flagged for IPD.

Item R_2ARIT0547B became easier in years 2006 and 2007 and was flagged for significant NCDIF values for those two years. The *b* parameter estimates were 0.42 and 0.41 for year 2006 and 2007, respectively, compared to 0.12 from 2004, and the differences in difficulty was most evident for theta levels below 0. The item was also easier in 2005 compared to 2004, but the differences were not large enough for the item to be flagged as drifting in 2005.

We looked into the content areas measured by the two items in order to understand the possible cause for the drift. Item M-070445 measures two subcontent areas: (1) subtraction and multiplication and (2) application and problem solving and, specifically, rate problems including ratio and proportion. Item R_2ARIT0547B also measures application and problem solving, focusing on percent problems. Neither of the two items was overexposed in each year of 2004 through 2007. Further research is needed to understand if there are changes in curriculum emphasis related to content measured by these two items, or if there are other possible explanations for the observed changes in item parameters.

Discussion

The current study aimed to empirically estimate item parameter drift at item level for the ACCUPLACER Arithmetic assessment. The results suggest that the Arithmetic test maintained a reasonably stable scale in the years 2004 through 2007.

The application of NCDIF to this particular adaptive test resulted in very few items being flagged using the criteria obtained from replications based on the 2004 data. The results do make sense in that the ICCs from the four years of calibrations tended to be very consistent from year to year, and the items that had the larger NCDIF were the ones evidencing the larger gaps in ICC. The results suggest that the possible concern about overexposure of certain Arithmetic items can be relieved because no item was identified as drifting due to overexposure. This is not unexpected. The placement assessments are considered low-stakes tests, and the results are used to place prospective students in the correct course level. Memorizing items would be counterintuitive as very few students would want to be placed into courses with a high probability of failure. On the other hand, the Arithmetic curriculum and teaching methods have remained relatively stable over the years. Therefore it is reasonable to observe few items showing parameter drift.

One caveat about this study was that the criteria used to determine the null distributions of the observed NCDIF values for each item were based on replicated samples of 15,000 drawn from the 2004 data. The NCDIFs for the yearly calibrations, on the other hand, were based on the entire test-taking populations from 2004 through 2007. The smaller samples used in the replications had likely introduced additional sampling error, making the NCDIF criteria conservative with regard to identifying outliers.

This study provided exploratory results for applying the NCDIF index to examine IPD in the CAT environment. Further item drift studies have been planned for other tests in the assessment battery. In respect to this study, the results were reasonable and reassuring, and the use of the NCDIF appears to be practical. Further research, particularly simulation research, is needed to determine variables that may impact the power of the NCDIF for detecting IPD with CAT and to determine how to set optimal critical values. In addition, simulation research is needed to examine what degree of IPD may cause errors in the linking process for scaling the item pools and lead to measurement bias in ability estimates.

References

Baker, F. B. (1992). *Item response theory*. New York: Springer-Verlag.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–449.

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

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on psychometrics of a cognitive ability test battery. *Journal of Applied Psychology*, *84*, 610–619.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265–300.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *20*, 369–377.

Guo, F., & Wang, L. (2005). Evaluating scale stability of a computer adaptive testing system. *GMAC Research Report*, RR-05-12.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Holland, P. W., & Wainer, H. (1993). *Differential Item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 51–66.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368.

Rupp, A. A., & Zumbo, B. D. (2003a, April). *Bias coefficients for lack of invariance in unidimensional IRT models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Rupp, A. A., & Zumbo, B. D. (2003b). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research*, *XLIX*, 264–276.

Stocking, M. L. (1988). Scale drift in online-calibration. *ETS Research Report*, RR-88-28-ONR.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77–87.

Wise, S. L., & Kingsbury, G. (2000). Practical Issues in developing and maintaining a computerized adaptive testing program. *Psicológica* (2000) *21*, 135–155.

Wollack, J. A., Sung, H. J., & Kang, T. (2006). *The impact of compounding item parameter drift on ability estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Appendix

Table A

Summary Statistics of Item Exposure Rates Across Years

Variable	*N*	*Mean*	*Std Dev*	*Min*	*Max*
expRate_04	223	0.0703	0.0785	0.0003	0.5519
expRate_05	223	0.0739	0.0804	0.0003	0.5534
expRate_06	223	0.0742	0.0807	0.0003	0.5529
expRate_07	223	0.0740	0.0798	0.0003	0.5439


CollegeBoard
inspiring minds™