

Comparative Study of Equating Methodology Versus Pre-Equated Panel Scoring

Technical Report

September 2007
Number W0701

Chuah Siang Chee

American Institute of CPAs



American Institute of Certified Public Accountants

American Institute of CPAs Comparative Study of Equating Methodology Versus Pre-Equated Panel Scoring

Introduction

The purpose of this study was to determine whether linear equating of panels leads to differences in scoring when compared with Item Response Theory (IRT) scoring. The results were evaluated based on differences in the pass/fail classification of examinees using the alternative equating method. The operational IRT score decisions from an actual test administration were treated as “true” for this study.

Operational scoring decisions for the Uniform CPA Examination were calculated using Item Response Theory (IRT) parameters. IRT parameters are considered to be population invariant and therefore the same parameters can be applied to different samples and the estimated abilities, or theta hats, for multiple choice questions and simulations are directly comparable i.e. they are on the same metric. Therefore, once item parameters are estimated in joint calibration of all items for panels, no further equating is considered necessary in order to report test scores across different test panels. Theta hat (using EAP) scores for candidates are converted to estimated number right scores and transformed to a normative scale before being reported to examinees.

On the other hand, linear equating uses a different methodology. One linear equating procedure is the mean-sigma procedure. Let us assume there are two test panels Panel X and Panel Y. In order to transform scores on Panel X to be equivalent to scores on Panel Y with linear equating, we would need a linear equation in the form of $l_y(x) = slope(x) + intercept$ with

$$\text{slope} = \frac{\sigma(\theta_Y)}{\sigma(\theta_X)}, \text{ and intercept} = \mu(\theta_Y) - \frac{\sigma(\theta_Y)}{\sigma(\theta_X)} \mu(\theta_X) \quad (1)$$

In the case of equating test panels for the Uniform CPA Examination which is a multistage adaptive examination, some test panels have overlapping sets of items (i.e. testlets). These common testlets can be used to equate test panels using the standard deviations and means of the IRT-scored testlets to compute the linear equation. The \bar{x} and σ from estimated number right scores are the basis of the equating slope and intercept.

Methods

Each portion of the Uniform CPA Examination consists of an adaptive multiple choice section, two simulation tasks, and constructed responses. The exception to this configuration is the business environment and concepts portion of the examination (BEC). The BEC does not have a simulation section or CRs and the multiple choice section is non-adaptive (i.e. linear format).

Two pairs of panels were selected from each of the test sections: AUD, FAR, BEC and REQ. Samples were selected from the 2005 Q1 window of the examination for analysis. Only samples with more than 300 examinees per test panel were selected for analysis. Missing data from the multiple choice and simulation section was treated as incorrect responses. Additionally, if an examinee fails to complete one of the two simulation tasks, they were also excluded from the analysis. Only operational items were used in the linking. Pre-test items were not included in any of the analysis. Table 1 lists the number of multiple-choice items in the equating testlets and sample sizes for each panel selected.

One requirement for linear equating is for samples of each test to be equivalent. That is to say, the distribution of abilities for examinees taking the test should be roughly equivalent. For the purposes of this study examinees taking each test panel are believed to be equivalent because panels are randomly assigned to candidates at the test center. However, the adaptive nature of the test leads to range restrictions for the second and third stage testlet samples. Candidates who are routed to an easier testlet cannot be considered equivalent to candidates routed to a hard testlet because this routing decision is based on examinee ability. Only testlets administered at the first phase of the panel have comparable distributions of examinee abilities. The linear equation is then applied to estimated number right scores from simulations and multiple choice scores to derive “equated” total scores.

The equating parameters were computed using Estimated Number Right (ENR) because overall scores are reported using ENR and not the observed number right. ENR is computed by converting theta hat scores using a conversion table. The multiple-choice section is scored using the theta metric because the multiple-choice test is adaptive. The adaptive nature results in the administration of items with different difficulties based of ability. Consequently, the number right score are not comparable for groups that receive different test items. The theta score can account for these differences, and therefore the theta scores of examinees that receive different items are directly comparable in an adaptive test.

The mean-sigma procedure was used to equate estimated number right scores for each pair of panels. The equating procedure was applied to only one panel (which we will call the target panel) in each pair so that test scores from that panel would be on the same metric as the other panel in the pair (which we will call the anchor panel). The ‘*t*’ designation designates the target panel and the ‘*a*’ designation designates the anchor panel in Table 1. Test scores from non-

pairs were not equated using the mean-sigma procedure. Therefore, comparisons between other panels were not included in the analysis.

Estimated number right scores from both the operational method and the linear equated method were then converted to the Raw Aggregate Score (*RAS*). The *RAS* is a weighted equation used to combine the three sections of the Uniform CPA Examination: simulation, multiple choice, and constructed response, into a single test score. Constructed responses were included in the scoring for completeness but are not transformed because the scores are not reported in the same metric ‘theta hat’ as the multiple choice scores.

$$RAS = (.2)l_y(x_{simulation}) + (.7)l_y(x_{multiple\ choice}) + (.1)Constructed\ Response \quad (2)$$

This score was then converted to the transformed aggregate score (*TAS*) where cut scores for the test have been established. The *RAS* score was then used to compute pass/fail classifications for analysis.

Results

Table 2 shows the average test scores for examinees. It consists of the estimated number right scores for examinees scored only with the equating panel, the average raw aggregate score, and the raw aggregate score transformed using the mean-sigma transformation. Only the target panels were transformed in order to match them to their matching anchor panels. Table 3 lists the slope and intercept parameters used for the mean-sigma transformation.

The differences in classification between IRT equating and mean-sigma equating is listed in Table 4. The panels for BEC and panel FAR14_p1828_28 from FAR appear to have the largest difference in classification between the IRT and mean-sigma equating procedures. The number of discrepancies in classification was highest for BEC in comparison to the other sections of the test. The percentage of classification discrepancies for the two pairs of BEC panels was 3.68% and 7.89%. This is higher than the any of the other panels from the other sections. Panel BEC01_p2246_46 in particular has the largest differences in pass/fail classification. The mean-sigma procedure resulted in 32 examinees from a sample of 401 being classified as passing, whereas the IRT procedure classifies those examinees as failing.

Discussion

The results suggest that there are no significant differences between the IRT and mean-sigma procedures. While some of the difference in pass/fail classification might appear to be large, the actual differences found between the IRT equating and mean-sigma equating are relatively small because of the distribution of examinee abilities. Many of the examinees have abilities that congregate around the pass/fail cut score. For example, panel BEC01_p2246_46 had a standard error of 3.39. There were 100 examinees, or 24.9% of the examinees within 1 standard error of the cut score. Having large numbers of examinees congregated around the cut score inflates the minor differences in scoring based on the two equating procedures.

Some of the discrepancy in classification found can be explained by small differences in the ability distributions between pairs of panels. The mean-sigma approach relies on having equivalent distributions of examinee abilities across both panels. However, we can see in Table

2, the average estimated number right scores for the equating testlets is not perfectly matched. These small differences in examinee abilities lead to differences in equating using the mean-sigma procedure. These small differences were then magnified by the congregation of examinee abilities around the cut score.

Part of the discrepancy in examinee ability distributions across testlets could be explained by errors in measurement. While the random assignment of test panels to examinees should have produced equivalent distributions of abilities across panels, the limited number of items in the equating testlets (20~25) reduces measurement precision. These differences in estimation partly explain small the differences in estimating ability for the equating testlets.

The BEC is the only section of the Uniform CPA Examination that is not adaptive and does not include a simulation or constructed response component. There was some concern that the simulation items and the adaptive component of the multiple-choice items might impact the equating procedure. If the BEC produced fewer differences in classification between the IRT and sigma-mean equating, it would suggest that the simulation items and adaptive component of the test be having an undesired effect on the equating methodology. However, the results proved contrary. The BEC produced more differences than the other sections. Therefore, the impact of the simulation items and adaptive component on the equating procedure does not appear to be a significant concern.

In summary, there appears to be no significant differences between the pre-equated IRT panel scoring and mean-sigma equating procedures. However due to the complexities of implementing the mean-sigma equating procedure in an adaptive test, the IRT methodology appears to be the most practical.

Table 1
 Panel Description For Multiple-Choice Items

Panel	# Examinees (N)	# Items In Equating Panel
AUD06_p0243_02 (a)	399	25
AUD06_p0243_43 (t)	440	25
AUD11_p0640_06 (a)	405	25
AUD11_p0640_40 (t)	397	25
BEC03_p0143_01 (a)	449	20
BEC03_p0143_43 (t)	435	20
BEC01_p2246_22 (a)	457	23
BEC01_p2246_46 (t)	401	23
FAR03_p0840_08 (a)	501	25
FAR03_p0840_40 (t)	496	25
FAR14_p1828_18 (a)	516	25
FAR14_p1828_28 (t)	507	25
REG01_p1128_11 (a)	491	20
REG01_p1128_28 (t)	466	20
REG03_p3345_33 (a)	497	20
REG03_p3345_45 (t)	462	20

(a) anchor panel, (t) target panel

ENR. Estimated Number Right for multiple-choice items only

Table 2
Average Examinee Test Scores

Panel	Average <i>ENR</i> for Examinees Scored Only With Equating Testlet	Average <i>RAS</i>	Average <i>RAS</i> After Mean-Sigma Transformation
AUD06_p0243_02 (a)	68.37	70.00	-
AUD06_p0243_43 (t)	67.68	70.00	70.54
AUD11_p0640_06 (a)	72.80	70.45	-
AUD11_p0640_40 (t)	73.20	71.64	71.23
BEC03_p0143_01 (a)	61.88	65.05	-
BEC03_p0143_43 (t)	59.69	64.68	63.67
BEC01_p2246_22 (a)	59.71	65.47	-
BEC01_p2246_46 (t)	60.59	64.47	66.58
FAR03_p0840_08 (a)	60.87	59.65	-
FAR03_p0840_40 (t)	60.83	60.56	60.61
FAR14_p1828_18 (a)	62.11	62.03	-
FAR14_p1828_28 (t)	60.86	58.95	60.13
REG01_p1128_11 (a)	63.85	64.69	-
REG01_p1128_28 (t)	64.12	65.63	65.44
REG03_p3345_33 (a)	66.43	65.44	-
REG03_p3345_45 (t)	66.05	65.18	65.56

ENR. Estimated Number Right for multiple-choice items only

RAS. Raw Aggregate Score computed by combining multiple-choice, simulation and constructed response scores.

BEC only consists of multiple-choice items.

Table 3
Mean-Sigma Transformation Parameters

Target Panels	Slope (A)	Intercept (B)
AUD06_p0243_43	0.970	2.725
AUD11_p0640_40	1.086	-6.695
BEC03_p0143_43	0.970	0.947
BEC01_p2246_46	0.982	3.257
FAR03_p0840_40	1.060	-3.638
FAR14_p1828_28	0.939	4.952
REG01_p1128_28	1.030	-2.197
REG03_p3345_45	0.914	6.065

Table 4

Classification Differences Between Mean-Sigma and IRT Pre-equated Panels

Target Panels	# False Positive	# False Negative	# Classification Difference	Classification Difference Rate
AUD06_p0243_43	5	-	5	1.14%
AUD11_p0640_40	-	2	2	0.50%
BEC03_p0143_43	-	16	16	3.68%
BEC01_p2246_46	32	-	32	7.98%
FAR03_p0840_40	5	-	5	1.01%
FAR14_p1828_28	12	-	12	2.37%
REG01_p1128_28	-	4	4	0.86%
REG03_p3345_45	1	6	7	1.52%