Systematic Equating Error with the Randomly-Equivalent Groups Design

An Examination of the Equal Ability Distribution Assumption

Per-Erik Lyrén
Ronald K. Hambleton

EM NO 61, 2008
ISSN 1103-2685

Abstract

In this study the equal ability distribution assumption associated with the randomly-equivalent groups equating design was investigated in the context of a selection test for admission to higher education. Test-takers' scores on anchor items from two subtests were estimated using information about test-taker performance on the regular subtests. The results showed that the estimated anchor test scores varied sufficiently so that the equal ability distribution assumption could be questioned. Consequently, we call for more caution when applying the randomly-equivalent groups design in the equating of tests. Equal ability groups is a convenient assumption to make but it can lead to systematic bias in the equating of test scores and this study provides a demonstration of that point.

Keywords: equating error, randomly-equivalent groups design, anchor tests, college admission tests.

Introduction

Any test used for selection to higher education has certain goals to reach and so certain requirements need to be met. Scores from these tests are used to aid in the decisions about which students have the best chance of performing well at a higher educational level. Problems arise however when scores from multiple versions of the test are used interchangeably as is the case when candidates are given up to five to ten years to use their admission test scores. In this situation, scores from the multiple forms need to be adjusted to be made comparable and interchangeable. This is accomplished with the well-known statistical procedure called "equating" (Kolen & Brennan, 2004). This means that in every case where test scores are valid for more than a single administration, a fair and well functioning equating from administration to administration becomes an important part in the validation and interchangeability of the test scores over time.

One of the most common equating designs used across the world is the randomly-equivalent groups design (see e.g., Kolen & Brennan, 2004; Braun & Holland, 1982). This design is used in the equating of several large-scale assessment tests, for example, the ACT (ACT, 2007) the Armed Services Vocational Aptitude Battery (ASVAB; Quenette, Nicewander, & Thomasson, 2006) and the General Ability Test (GAT; National Assessment and Examinations Center, 2005), and it was used in the equating of the Medical College Admission Test (MCAT; Hendrickson & Kolen, 1999) until 2007 when they switched from paper-based administration to computer-based administration. It is also a popular design when field-testing large numbers of test items assigned to forms that are distributed to examinees on a more or less random basis. Its popularity is largely due to the ease of administration of test forms because each group has to take only one test form and it is not necessary to have common items in the test forms.

When using the randomly-equivalent groups design one assumes that any differences in actual test performance can be attributed to the use of non-parallel tests, and test scores are revised accordingly. Consequently, that equating design makes the strong assumption that the ability distributions from the test administrations are equivalent. But this assumption is very strong, and almost certainly false in countries where educational reform is taking place, and so the abilities of students are not likely to be consistent from one test administration to the next. More students may be seeking a college experience in some countries, students may be, on the average, better over time because of educational reform, etc.

There are many reasons to believe that the assumption of equal ability of candidates over test administrations is false, and when it is, it will have consequences for test developers as well as test-takers. For instance, due to year-to-year changes in the pool of candidates, those responsible for the test used for admission to the Faculty of Social Sciences at the University of Tartu in Estonia and the Estonian National Defense College have decided to let these test scores be valid for only one administration and thereby avoid having to equate at all (O. Must, personal communication, March 7, 2008). Paradoxically, when students in a group do quite well, it is assumed that the test is relatively easy and test scores through the equating process are lowered. If the test was not easier, then the better students, the ones the universities are most interested in, have their scores lowered and they cannot be distinguished from lower performing students who were administered other forms.

The purpose of this study was to examine the equal ability distribution assumption underlying the randomly-equivalent groups equating design using data from the Swedish Scholastic Assessment Test (SweSAT; Stage & Ögren, 2004), and to discuss the consequences on the outcome of the equating when that assumption is violated. The SweSAT is, next to the upper-secondary school GPA, the most widely used instrument for selection to higher education in Sweden. To facilitate this study there are potential linking items available that originate from experimental studies on the SweSAT. These items have not been intended for equating studies, and due to lack of content coverage, they would not be suitable for use in a common-item equating design. However, they will be used in this study to highlight the problem of the randomly-equivalent groups design.

There are several reasons why questions have been raised about the appropriateness of using the current equating design on the SweSAT. First, the size and composition of the pool of test-takers have changed. The number of people taking the SweSAT has decreased about 45 percent from 1997 to 2006. During the same period the proportion of test-takers made up by upper-secondary school seniors who attend an academically-oriented program have decreased from 27 percent to 15 percent. Both these facts suggest that something is happening with the pool of test-takers. Second, the upper-secondary school in Sweden has undergone many changes during the last few years in terms of re-organizations and syllabi makeovers. There are reports of students leaving upper-secondary school with insufficient knowledge, especially in mathematics, to manage higher education (Skolverket, 2005; Carlsson, 2002; Nilsson, 2003). These facts suggest that the main assumption of group ability equivalence underlying the randomly-equivalent groups design could be violated, and thereby causing a major threat to the validity of college admissions scores. We suspect that the problem in Sweden with the SweSAT is not unusual across Europe and the rest of the world. Besides, from these specific issues the importance of checking the validity of the equal ability distribution assumption is highlighted in Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), where it is stated that "In equating studies that rely on the statistical equivalence of examinee groups receiving different forms, methods of assuring such equivalence should be described in detail" (see Standard 4.12, p. 58).

The SweSAT

The SweSAT is a norm-referenced, paper-and-pencil test that has been used for selection to higher education since 1977. The construct of the test is based on general developed abilities that one needs to perform well at the college and university level, such as reading comprehension and quantitative ability. The construct is measured through five subtests: Vocabulary (WORD), Swedish Reading Comprehension (READ), English Reading Comprehension (ERC), Data Sufficiency (DS), and Diagrams, Tables, and Maps (DTM). However, only the composite scale scores, which range from 0.0 to 2.0 with 0.1 increments, are used in the selection process.

The test is administered twice a year, and takes a full day of testing time. The test day is divided into five 50-minute blocks, of which WORD and ERC together make up one block, and READ, DS and DTM make up one block each. The fifth block is for try-out items, and this block is made up of one of the previously mentioned constellations (WORD + ERC; READ; DS; DTM). What subtest that makes up the try-out block varies between different administrative regions; some regions will have WORD and ERC in the try-out block, while others will have READ, DS, or DTM. The order of the blocks during the test day changes from one administration to another, and the test-takers do not know the order of the blocks in advance.

This means that the item statistics for the try-out items are reliable, and thus it is possible to make well-founded decisions about the quality of the try-out items. On the other hand, the design in place does not guarantee that the new items are administered to a representative sample of examinees.

The SweSAT equating procedure

In order that the ability distribution of the total group not be relied on too heavily in the equating procedure, two reference groups are used with each test administration. Reference group I (RG I) is selected through a proportionally stratified selection, such that the distribution of sex, age and educational background is the same across test administrations. Reference group II (RG II) consists of seniors in upper-secondary school (18 or 19 years old) attending an academically oriented program. Together with the Total Group (i.e., all test-takers from one administration) they make up the three equating groups. The equating procedure is carried out separately in all three equating groups, and the final equating function is a weighted mean of the three individual equating functions. When deciding on the final equating function, the Total Groups solution has the largest weight.

The pivotal part of the SweSAT equating procedure is to define the limits of the *j*'th raw score interval associated with a scale score s_j on the new test (the test to be equated) that corresponds to the same proportion of respondents with a raw score in that interval as the *j*'th interval on the reference tests (Emons, 1998). Note the term "reference test*s*": Because test scores are valid for five years it is necessary to equate the new test not only to the last test, but to all tests during the preceding five years.

The equating procedure can be defined formally in the following way: Let raw scores be denoted by r, $r \in (0, 1, ..., 122)$, and let scale scores be denoted by s, $s \in (0.0, 0.1, ..., 2.0)$. Furthermore, let the upper and lower limits of the raw score interval assigned to a scale score s be denoted by r_{ls} and r_{us}, respectively, such that test-takers with a raw score in the interval (r_{ls}, r_{us}) are assigned a scale score s. Then, let the proportion of test-takers obtaining a scale score s be denoted by $p(s)$, and let the raw score frequency distribution from test g and any scale score frequency distribution be denoted by $p_g(r)$ and $q(s)$, respectively. The cumulative frequency distributions are given by

$$P_g(r) = \sum_{r=0}^{122} p_g(r), \qquad (1)$$

and

$$Q(s) = \sum_{s=0.0}^{2.0} q(s). \qquad (2)$$

The goal of the equating procedure is to define r_{ls} and r_{us} that yield $P_1(r_{us})$ in the new population similar to $P_2(r_{us})$, $P_3(r_{us})$, and so on, in the old (reference) populations. In other words, we strive for the following equality to hold: $P_1(r) = Q(s) = P_2(r) = P_3(r) = ... = P_g(r)$. The equating procedure starts with identifying the raw score interval assigned to the highest scale score, $s = 2.0$. Given the lower limit of the raw score interval associated with $s = 2.0$ from the reference tests, the next step is to find the percentage of test-takers with a score at or below that raw score limit and subsequently the raw score on the new test that has the same percentage of test-takers at or below that raw score. These scores are then considered to be equivalent. Furthermore, when the lower limit of the raw score interval assigned to the highest scale score ($r_{l2.0}$) has been identified, the upper limit of the adjacent interval ($r_{u1.9}$) has been identified as well (because, by definition, $r_{u1.9} = r_{l2.0} - 1$). Then the lower limit of the raw score interval assigned to the next scale score, $s = 1.9$, is determined in the same manner as for the highest scale score, and then the procedure is repeated for $s = 1.8, \dots, 0.0$. This equipercentile equating procedure results in equated raw score intervals, which are then converted to the score scale (Emons, 1998). An example of equating functions for the SweSAT is given in Table 1.

Table 1

Equating Functions for Three SweSAT Administrations

	2006, Spring		2005, Fall		2005, Spring	
Scale score	CRF	Raw score interval	CRF	Raw score interval	CRF	Raw score interval
0.0	2.7	0– 33	2.4	0– 33	2.4	0– 32
0.1	6.9	34– 39	6.7	34– 39	6.5	33– 38
0.2	10.0	40– 42	9.8	40– 42	9.7	39– 41
0.3	13.4	43– 45	13.2	43– 45	13.4	42– 44
0.4	17.5	46– 48	17.6	46– 48	17.8	45– 47
0.5	23.3	49– 52	22.3	49– 51	22.6	48– 50
0.6	30.0	53– 56	29.7	52– 55	29.3	51– 54
0.7	36.8	57– 60	37.4	56– 59	36.9	55– 58
0.8	43.9	61– 64	45.0	60– 63	44.6	59– 62
0.9	50.9	65– 68	52.5	64– 67	52.1	63– 66
1.0	58.2	69– 72	59.7	68– 71	59.3	67– 70
1.1	65.1	73– 76	66.6	72– 75	66.3	71– 74
1.2	71.7	77– 80	73.0	76– 79	72.7	75– 78
1.3	77.8	81– 84	78.8	80– 83	78.6	79– 82
1.4	84.3	85– 89	84.1	84– 87	84.8	83– 87
1.5	88.6	90– 93	88.6	88– 91	89.2	88– 91
1.6	93.3	94– 98	93.0	92– 96	93.4	92– 96
1.7	96.0	99–102	95.8	97–100	96.1	97–100
1.8	98.0	103–106	97.9	101–104	98.0	101–104
1.9	99.2	107–110	99.0	105–108	99.2	105–108
2.0	100.0	111–122	100.0	109–122	100.0	109–122

Note. CRF = cumulative relative frequency.

In Table 1, for each of the three administrations, each step on the score scale is given, with the associated cumulative relative frequency and raw score interval. For example, a raw score between 69 and 72 on the spring 2006 test is considered to be equal to a raw score between 68 and 71 on the fall 2005 test and equal to a raw score between 67 and 70 on the spring 2005 test. This is because the cumulative relative frequency distributions associated with those raw score intervals are approximately equal.

Method

Data

The data used in this study are item responses from the SweSAT subtests WORD and DS. There are 15 WORD anchor items (referred to as the WORD anchor test) administered in try-out sections in positions 1–5, 21–25, and 35–39 from the 1997 fall test to and including the 2004 spring test, and the 22 DS anchor items (referred to as the DS anchor test) constitute a full try-out section administered from the 1999 fall test and onwards. All anchor items have been piloted and selected to meet certain criteria, such as content relevance, difficulty, and discrimination. The number of items in the regular subtests and the anchor tests can be seen in Table 2, and intercorrelations between the subtests are found in Table 3.

Table 2

Number of Items in the SweSAT Regular Subtests and Anchor Tests

Subtest	Regular Test	Anchor Test
Vocabulary (WORD)	40	15
Swedish Reading Comprehension (READ)	20	N/A
English Reading Comprehension (ERC)	20	N/A
Data Sufficiency (DS)	22	22
Diagrams, Tables, and Maps (DTM)	20	N/A
Total	122	37

Table 3

Intercorrelations of the Five Subtests of the SweSAT

Subtest	WORD	DS	READ	DTM	ERC
WORD	-	.37	.63	.40	.60
DS		-	.49	.64	.49
READ			-	.50	.65
DTM				-	.48
ERC					-

Note. Correlations are the averages from five test administrations: 2004 Spring, 2004 Fall, 2005 Spring, 2005 Fall, and 2006 Spring.

Procedure

The equal ability assumption was examined by studying the mean performance on the two anchor tests in the three different equating groups (the Total Group, RG I, and RG II). Unfortunately, each anchor test is administered in only one of a number of administrative regions. Because the test-taking groups in different regions are not selected completely at random and therefore may differ quite considerably in terms of background variables, it is necessary to estimate the performance of the total group based on data in the try-out (anchor) group. The estimation can be done in several ways, but as the score distributions are approximately normal and only group level estimates are of interest, classical linear equating seems appropriate and was applied in this study.

First, the regular test and the anchor test were equated (in the try-out group) by setting the standardized deviation scores (*z*-scores) on the regular test and the anchor test to be equal, such that

$$\frac{r-\mu(R)}{\sigma(R)}=\frac{a-\mu(A)}{\sigma(A)}, \tag{3}$$

where *μ(R)*, *σ(R)*, *μ(A)*, and *σ(A)* are the mean and standard deviations of the regular test and the anchor test, respectively, and *r* and *a* are test-takers' scores on the regular test and the anchor test, respectively. Then, by using any test-taker's score on the regular test (*r*) and re-arranging Equation 3, one can estimate that test-taker's score on the anchor test ($\hat{a}$) through the following equation:

$$\hat{a} = \frac{\sigma(A)}{\sigma(R)}\left(r - \mu(R)\right) + \mu(A)\,. \qquad (4)$$

With $\hat{a}$s for all test-takers, the estimated mean and standard deviation of the scores on the anchor test can easily be computed by taking the mean and standard deviation of all $\hat{a}$s, that is

$$\hat{\mu}(A) = \mu(\hat{a})\,, \qquad (5)$$

and

$$\hat{\sigma}(A) = \sigma(\hat{a})\,. \qquad (6)$$

The comparison of performance within groups across test administrations is done in two ways. First, the estimated mean scores are compared to identify the largest score difference on the DS anchor test and the WORD anchor items over any five-year period. The motivation for studying five-year periods is that SweSAT scores are valid for college applications for five years. That score difference is then evaluated with regards to the concept of a *difference that matters* (Dorans & Feigenbaum, 1994; Liu, Cahn, & Dorans, 2006). In the context of the SweSAT, an unrounded raw-score difference of half a score unit or larger would be a difference that matters because it would lead to a different score conversion table. Second, the groups are compared with regards to the effect size of the difference of mean scores. The comparison is done between the three equating groups as well as within the three equating groups; for example, the Total Group is compared to RG I at each administration (i.e., between-groups comparison) and RG II at one administration is compared to RG II at another administration (i.e., within-groups comparison). The measure of effect size is Cohen's *d* (Cohen, 1988), that is

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}\,. \qquad (7)$$

By comparing the three equating groups with regards to effect size, it is possible to gain an understanding of the relative performance of the groups. This is important information because the final equating function is a result of the individual equating functions in the three groups. Therefore, if the relative performance of the equating groups fluctuates extensively, then that will affect the outcome of the equating.

Results

The results are presented first for the WORD anchor test and then for the DS anchor test. For each test, the score distributions are presented first and then the effect sizes.

The WORD Anchor Test

As can be seen in Table 4 and Figure 1, the largest mean score difference on the WORD anchor items is 0.80 in the Total Group, 1.18 in RG I and 1.04 in RG II. The corresponding effect sizes are 0.27 in the Total Group, 0.38 in RG I and 0.39 in RG II. An effect size of 0.39 is large in this context because that means there is about a 25 percent non-overlap of the two distributions, and that is clearly an indication that the two distributions cannot be considered equal. If transformed to the WORD regular raw-score scale, by multiplying scores by 40/15, the expected difference would be 2.13 in the Total Group, 3.15 in RG I, and 2.77 in RG II . These are clearly differences that matter; in fact, a rounded raw-score difference of 3 points would result in a different equating function at the majority of the raw-score levels. Also, the mean score for RG II at four of the last five administrations are the lowest ones for that group during the whole period.

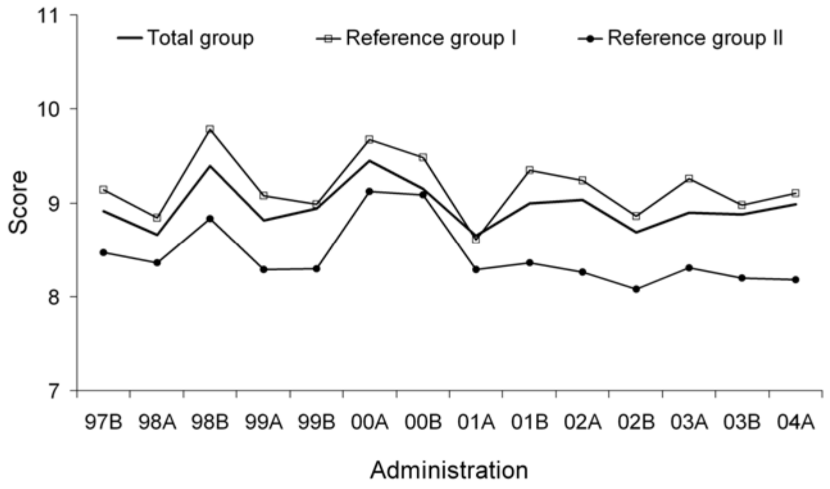

Figure 1 Mean scores on the WORD anchor test for the three equating groups. (A are spring administrations; B are fall administrations)

Table 4

Means and Standard Deviations of Scores on the WORD Anchor Items in the Three Equating Groups

Test	Total Group		RG I		RG II	
	M	SD	M	SD	M	SD
1997 Fall	8.91	2.97	9.14	3.39	8.47	2.76
1998 Spring	8.66	2.97	8.84	3.24	8.36	2.81
1998 Fall	9.40	2.87	9.79	3.08	8.83	2.53
1999 Spring	8.81	3.06	9.08	3.35	8.29	2.77
1999 Fall	8.94	2.92	8.99	3.30	8.30	2.64
2000 Spring	9.45	2.93	9.68	3.23	9.12	2.72
2000 Fall	9.14	2.86	9.48	3.26	9.08	2.51
2001 Spring	8.65	2.97	8.61	3.18	8.29	2.93
2001 Fall	8.99	2.96	9.36	3.14	8.36	2.90
2002 Spring	9.02	3.03	9.24	3.29	8.26	2.75
2002 Fall	8.69	3.15	8.86	3.56	8.08	2.66
2003 Spring	8.90	2.87	9.26	2.97	8.30	2.62
2003 Fall	8.87	2.83	8.98	3.69	8.20	2.44
2004 Spring	8.99	2.96	9.10	3.22	8.18	2.47

When it comes to the effect size of differences in mean scores between the different equating groups (Figure 2), one interesting observation can be made. The effect size between the Total Group and RG II increases quite consistently from basically 0 on the fall 2000 form to 0.30 on the spring 2004 form; this indicates that the performance of RG II relative to the Total

Group has decreased. This is probably largely due to the fact that the size of RG II relative to the Total Group has decreased.

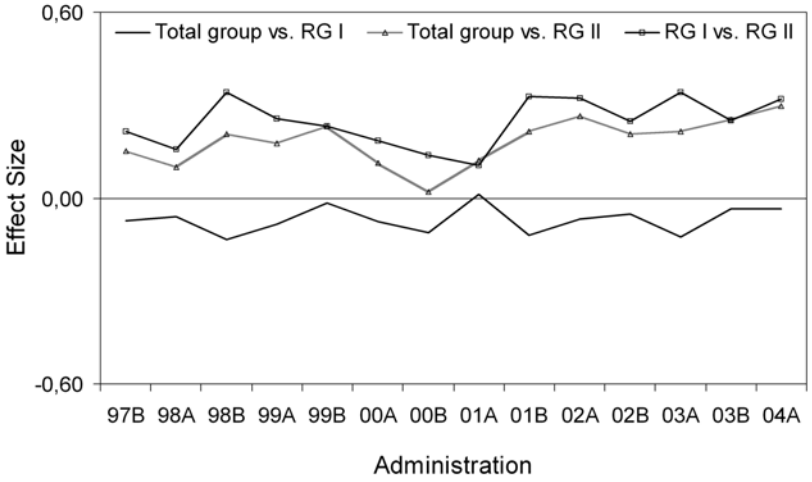

Figure 2 Effect size of the difference in scores on the WORD anchor test between the three equating groups

The DS Anchor Test

The distributions of scores on the DS anchor test are displayed in Table 5 and Figure 3. However, before going any deeper into the result pattern for that test one important note needs to be made. The distinct peaks on the fall 2001 and fall 2004 forms can probably be attributed to the effects of test-wiseness (e.g., see Rogers & Bateson, 1991). In this case, the test-wiseness effects seem largely due to the order in which the subtests were administered, where on the fall 2001 form the DS anchor test was administered directly after the regular DS test, and on the fall 2004 form the DS anchor test was administered in Block 3 and the regular DS test in Block 1 (i.e., there was another test in Block 2, between the two DS tests). It is hypothesized that during the first administration of the test the test-takers learned the format and time management strategies, which had a positive effect on test taker performance on the next administration of the DS test. This means that the estimated scores on the anchor test on those occasions were inflated, and therefore they cannot be assumed to represent the true performance of the groups. A recent study (Lexelius, 2007) supports the hypothesis that the order of the subtests had an effect on the performance of test-takers.

Table 5

Means and Standard Deviations of Scores on the DS Anchor Test in the Three Equating Groups

Test	Total Group		RG I		RG II	
	M	SD	M	SD	M	SD
1999 Fall	11.79	5.11	10.10	5.33	12.90	4.72
2000 Spring	12.03	5.01	10.23	5.21	13.04	4.85
2000 Fall	11.97	4.85	10.41	5.15	13.01	4.73
2001 Spring	11.77	5.10	10.04	4.88	13.06	4.91
2001 Fall	12.81	5.11	11.03	5.25	14.01	4.33
2002 Spring	11.42	5.11	9.89	5.20	12.63	4.56
2002 Fall	11.70	5.12	10.00	5.25	12.25	4.42
2003 Spring	11.84	4.94	10.18	5.33	12.30	4.27
2003 Fall	11.89	5.09	10.05	5.12	12.86	4.63
2004 Spring	11.47	4.92	10.27	5.28	12.16	4.41
2004 Fall	12.49	5.04	10.57	4.66	13.19	4.93
2005 Spring	11.99	5.04	10.52	5.07	12.57	4.79
2005 Fall	11.46	5.09	10.01	4.48	12.11	5.12
2006 Spring	11.70	5.06	10.49	5.14	12.39	4.57

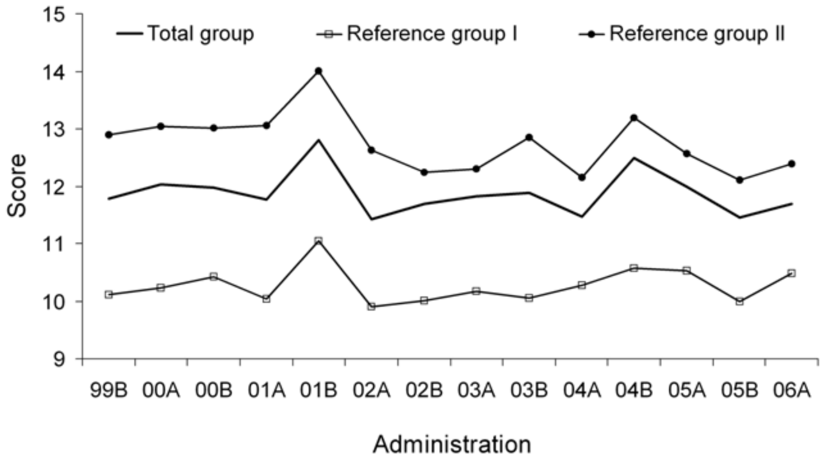

Figure 3 Mean scores on the DS anchor test for the three equating groups

Disregarding the results on the fall 2001 and fall 2004 forms (which seems to be the most defensible decision), the largest mean score difference on the DS anchor test was 0.61 in the Total Group, 0.63 in RG I and 0.93 in RG II, while the largest effect size was 0.11 in the Total Group, 0.12 in RG I and 0.19 in RG II. However, if results from fall 2001 and fall 2004 are included, the largest mean score difference is 1.39 in the Total Group, 1.14 in RG I and 1.90 in RG II, and the largest effect size is 0.27 in the Total Group, 0.22 in RG I and 0.42 in RG II. Either way, it seems clear that scores for RG II fluctuated through the whole period, and the general trend seems to be going somewhat downward. Moreover, all the previously mentioned score differences are above the limit of what can be considered a score difference that matters, i.e. half a raw-score unit.

The effect size of the difference in mean scores on the DS test between the different equating groups (Figure 4) does not show any striking results. The effect size between the Total Group and RG II was quite consistent over the period, which in part can be explained by the relatively large size of RG II. However, the pattern of effect sizes between RG I and RG II shows an upward trend, which implies that the ability of RG II relative to RG I has been declining.

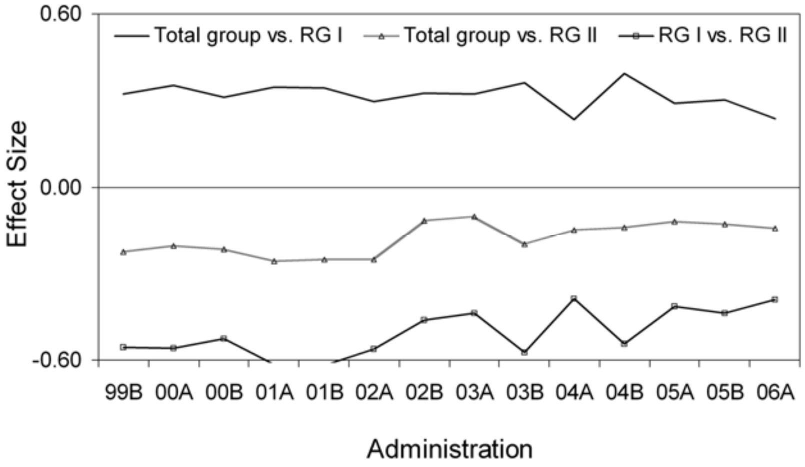

Figure 4 Effect size of the difference in scores on the DS anchor test between the three equating groups

Discussion

The purpose of this study was to examine the equal ability distribution assumption underlying the randomly-equivalent groups design used when equating the SweSAT, and to discuss the impact any violation of that assumption would have on the outcome of the equating. The results indicate that the score distributions on both the WORD anchor items and the DS anchor items were not stable over time in any of the three equating groups. The implications of this for the equating procedure is difficult to assess; however, what is clear is that if the composition of the test-taking group continues to change as much as it has done over the last few years, results from the current randomly-equivalent groups equating design will become even more problematic.

An important issue to consider is whether the results can be generalized from the two examined subtests to the whole test. From the intercorrelations in Table 3, it is clear that WORD is highly correlated with the other two verbal subtests (READ and ERC) and DS is highly correlated with the other quantitative subtest (DTM). This implies that the differences found would probably be reinforced rather than cancelled out by the other subtests. However, this does not mean that the WORD and DS anchor items can be used in an operational equating because, as pointed out previously, an operational anchor test should be a mini-version of the test in terms of the measured construct and content coverage.

The most serious concern from the study is the unstable performance of RG II. To deal with this, it is necessary to reduce the weight of that reference group in the equating procedure. A concrete consequence of this issue is that seniors in upper-secondary school might be advantaged against their peers in previous cohorts. If so, that is not the only way they are being advantaged when applying to higher education. For example, it has been shown that the upper-secondary school grades have been subject to quite severe inflation since the late 1990s (Wikström & Wikström, 2005; Cliffordson, 2004). This means that applicants with recent upper-secondary school GPAs have a better chance of being admitted on the basis of their GPA than applicants who have the same expected academic performance but who graduated from upper-secondary school some years ago.

An important issue is whether the population of test-takers is expected to be improving or not. From the view of those responsible for the SweSAT testing program the test-takers are not expected to be improving; if they did, there would be no justification for applying the current equating procedure. Also, even though there is no strong accountability system in Sweden, such as the No Child Left Behind Act in USA, from an educational policy point of view there are perhaps not expectations but at least hopes that a cohort leaving upper-secondary school and entering college a certain year would have a higher level of knowledge than previous cohorts. The performance of the population of test-takers is also highly dependent on its composition, and it is more than likely that the population will change due to changing demands for entering higher education. For example, it can be expected that because of the current decrease in applications to higher education that has no concomitant decrease in available study places, there will be no need for selection among candidates at many of the programs. As a result fewer people will feel a need to take the SweSAT and this will inevitably lead to a change in the population of test-takers. How this affects the performance of the population of test-takers remains to be seen.

Further research on this topic would be to investigate means for improving the SweSAT equating procedure. A natural start would be to examine the possibility of applying an anchor design much like is done in the US with the Scholastic Assessment Test (SAT); more specifically if it is feasible to use try-out items in the equating procedure. The most important implication of an improved equating procedure would be an increase in the validity of test scores (Kane, 2006; Messick, 1989). The construct validity would increase because one source of systematic error variance would be eliminated, which would lead to test scores being more useful for predicting academic performance. Also, test scores would be interpreted as a more precise measure that is useful for selection to higher education, and any possible negative consequences originating from the equating procedure would decrease. Test scores that are fairer and more comparable across test administrations would have positive effects for individuals as well as academic institutions and the admission system as a whole. Individual test-takers would feel they are being assessed more fairly, institutions would probably get students who are better prepared for higher education, and the credibility of the admission system as a whole would increase.

To conclude, the equal ability assumption underlying the randomly-equivalent groups design in the equating of the SweSAT is not severely violated at this time. However, the continuing change in the composition of the test-taking group calls for a need to apply an equating design that is not dependent on an untestable and questionable assumption, however convenient it may be to apply in practice. Other test agencies should be informed by this study too that the validity of test equating can be seriously threatened when intact groups form the basis for the equal ability group assumption.

References

ACT (2007). *ACT technical manual.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Braun, H., & Holland, P. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. Holland & D. Rubin (Eds.), *Test equating*. New York: Academic Press.

Carlsson, H., Weibull, T., Järner, S., Blomqvist, H., Franklin, H., Tranberg, H. et al. (2002, June 7). Studenterna blir alltmer okunniga [The students are becoming less and less knowledgeable]. *Dagens Nyheter.* Retrieved September 25, 2006, from www.dn.se.

Cliffordson, C. (2004). Betygsinflation i de målrelaterade betygen [Inflation in goal-related grades from upper-secondary schools]. *Pedagogisk Forskning i Sverige, 9*(1), 1–14.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS RM-94-10). Princeton, NJ: Educational Testing Service.

Emons, W. H. M. (1998). *Nonequivalent groups IRT observed score equating: Its applicability and appropriateness for the Swedish Scholastic Aptitude Test* (EM No. 32). Umeå, Sweden: Umeå University, Department of Educational Measurement.

Hendrickson, A. B., & Kolen, M. J. (1999). *IRT equating of the MCAT* (MCAT Monographs). Washington, DC: Association of American Medical Colleges.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4 ed., pp. 17–64). Westport, CT: American Council on Education/Praeger Publishers.

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

Lexelius, A. (2007). *Sekventiella effekter i högskoleprovet avseende delprovet NOG* [Sequential effects in the SweSAT subtest DS] (BVM No. 29). Umeå, Sweden: Umeå University, Department of Educational Measurement.

Liu, J., Cahn, M. F., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linkage of New SAT to old SAT across gender groups. *Journal of Educational Measurement, 43*(2), 113–129.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Educational Measurement/Macmillan.

National Assessment and Examinations Center (2005). *Unified national university entry examinations 2005* (Report). Tblisi, Georgia: Author. Retreived January 10, 2008, from http://naec.ge/files/433_ENG-Report-on-2005-University-entry-examinations.pdf

Nilsson, H. (2003). *Nya studenters kunskaper i matematik. En studie av matematikkunskaperna hos nya ingenjörsstudenter vid Växjö universitet 2002* [New students' knowledge in mathematics. A study of mathematical knowledge among new engineering students at Växjö University] (Rapporter från MSI No. 03047). Växjö, Sweden: Växjö University.

Quenette, M. A., Nicewander, W. A., & Thomasson, G. L. (2006). Model-based versus empirical equating of test forms. *Applied Psychological Measurement, 30*(3), 167–182.

Rogers, T. W., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*(2), 159–183.

Skolverket (2005). *Gymnasieutbildade är inte tillräckligt förberedda för högre studier* [Upper-secondary school leavers are not sufficiently prepared for higher education; Press release]. Retrieved September 22, 2006, from http://www.skolverket.se/sb/d/204/a/5057

Stage, C., & Ögren, G. (2004). *The Swedish Scholastic Assessment Test (SweSAT): Development, results and experiences* (EM No. 49). Umeå, Sweden: Umeå University, Department of Educational Measurement.

Wikström, C., & Wikström, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review, 24*, 309–322.

EDUCATIONAL MEASUREMENT

Reports already published in the series

EM No 1. SELECTION TO HIGHER EDUCATION IN SWEDEN. Ingemar Wedman

EM No 2. PREDICTION OF ACADEMIC SUCCESS IN A PERSPECTIVE OF CRITERION-RELATED AND CONSTRUCT VALIDITY. Widar Henriksson, Ingemar Wedman

EM No 3. ITEM BIAS WITH RESPECT TO GENDER INTERPRETED IN THE LIGHT OF PROBLEM-SOLVING STRATEGIES. Anita Wester

EM No 4. AVERAGE SCHOOL MARKS AND RESULTS ON THE SWESAT. Christina Stage

EM No 5. THE PROBLEM OF REPEATED TEST TAKING AND THE SweSAT. Widar Henriksson

EM No 6. COACHING FOR COMPLEX ITEM FORMATS IN THE SweSAT. Widar Henriksson

EM No 7. GENDER DIFFERENCES ON THE SweSAT. A Review of Studies since 1975. Christina Stage

EM No 8. EFFECTS OF REPEATED TEST TAKING ON THE SWEDISH SCHOLASTIC APTITUDE TEST (SweSAT). Widar Henriksson, Ingemar Wedman

1994

EM No 9. NOTES FROM THE FIRST INTERNATIONAL SweSAT CONFERENCE. May 23 - 25, 1993. Ingemar Wedman, Christina Stage

EM No 10. NOTES FROM THE SECOND INTERNATIONAL SweSAT CONFERENCE. New Orleans, April 2, 1994. Widar Henriksson, Sten Henrysson, Christina Stage, Ingemar Wedman and Anita Wester

EM No 11. USE OF ASSESSMENT OUTCOMES IN SELECTING CANDIDATES FOR SECONDARY AND TERTIARY EDUCATION: A COMPARISON. Christina Stage

EM No 12. GENDER DIFFERENCES IN TESTING. DIF analyses using the Mantel-Haenszel technique on three subtests in the Swedish SAT. Anita Wester

1995

EM No 13. REPEATED TEST TAKING AND THE SweSAT. Widar Henriksson

EM No 14. AMBITIONS AND ATTITUDES TOWARD STUDIES AND STUDY RESULTS. Interviews with students of the Business Administration study program in Umeå, Sweden. Anita Wester

EM No 15. EXPERIENCES WITH THE SWEDISH SCHOLASTIC APTITUDE TEST. Christina Stage

EM No 16. NOTES FROM THE THIRD INTERNATIONAL SweSAT CONFERENCE. Umeå, May 27-30, 1995. Christina Stage, Widar Henriksson

EM No 17. THE COMPLEXITY OF DATA SUFFICIENCY ITEMS. Widar Henriksson

EM No 18. STUDY SUCCESS IN HIGHER EDUCATION. A comparison of students admitted on the basis of GPA and SweSAT-scores with and without credits for work experience. Widar Henriksson, Simon Wolming

1996

EM No 19. AN ATTEMPT TO FIT IRT MODELS TO THE DS SUBTEST IN THE SweSAT. Christina Stage

EM No 20. NOTES FROM THE FOURTH INTERNATIONAL SweSAT CONFERENCE. New York, April 7, 1996. Christina Stage

1997

EM No 21. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SWESAT. A study of the DTM subtest. Christina Stage

EM No 22. ITEM FORMAT AND GENDER DIFFERENCES IN MATHEMATICS AND SCIENCE. A study on item format and gender differences in performance based on TIMSS´data. Anita Wester, Widar Henriksson

EM No 23. DO MALES AND FEMALES WITH IDENTICAL TEST SCORES SOLVE TEST ITEMS IN THE SAME WAY? Christina Stage

EM No 24. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the ERC Subtest. Christina Stage

EM No 25. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the READ Subtest. Christina Stage

EM No 26. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the WORD Subtest. Christina Stage

EM No 27. DIFFERENTIAL ITEM FUNCTIONING (DIF) IN RELATION TO ITEM CONTENT. A study of three subtests in the SweSAT with focus on gender. Anita Wester

EM No 28. NOTES FROM THE FIFTH INTERNATIONAL SWESAT CONFERENCE. Umeå, May 31 – June 2, 1997. Christina Stage

1998

EM No 29. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest WORD. Christina Stage

EM No 30. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest ERC. Christina Stage

EM No 31. NOTES FROM THE SIXTH INTERNATIONAL SWESAT CONFERENCE. San Diego, April 12, 1998. Christina Stage

1999

EM No 32. NONEQUIVALENT GROUPS IRT OBSERVED SCORE EQUATING. Its Applicability and Appropriateness for the Swedish Scholastic Aptitude Test. Wilco H.M. Emons

EM No 33. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest READ. Christina Stage

EM No 34. PREDICTING GENDER DIFFERENCES IN WORD ITEMS. A Comparison of Item Response Theory and Classical Test Theory. Christina Stage

EM No 35. NOTES FROM THE SEVENTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 3–5, 1999. Christina Stage

2000

EM No 36. TRENDS IN ASSESSMENT. Notes from the First International SweMaS Symposium Umeå, May 17, 2000. Jan-Olof Lindström (Ed)

EM No 37. NOTES FROM THE EIGHTH INTERNATIONAL SWESAT CONFERENCE. New Orleans, April 7, 2000. Christina Stage

2001

EM No 38. NOTES FROM THE SECOND INTERNATIONAL SWEMAS CONFERENCE, Umeå, May 15-16, 2001. Jan-Olof Lindström (Ed)

EM No 39. PERFORMANCE AND AUTHENTIC ASSESSMENT, REALISTIC AND REAL LIFE TASKS: A Conceptual Analysis of the Literature. Torulf Palm

EM No 40. NOTES FROM THE NINTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 4–6, 2001. Christina Stage

2002

EM No 41. THE EFFECTS OF REPEATED TEST TAKING IN RELATION TO THE TEST TAKER AND THE RULES FOR SELECTION TO HIGHER EDUCATION IN SWEDEN. Widar Henriksson, Birgitta Törnkvist

2003

EM No 42. CLASSICAL TEST THEORY OR ITEM RESPONSE THEORY: The Swedish Experience. Christina Stage

EM No 43. THE SWEDISH NATIONAL COURSE TESTS IN MATHEMATICS. Jan-Olof Lindström

EM No 44. CURRICULUM, DRIVER EDUCATION AND DRIVER TESTING. A comparative study of the driver education systems in some European countries. Henrik Jonsson, Anna Sundström, Widar Henriksson

2004

EM No 45. THE SWEDISH DRIVING-LICENSE TEST. A Summary of Studies from the Department of Educational Measurement, Umeå University. Widar Henriksson, Anna Sundström, Marie Wiberg

EM No 46. SweSAT REPEAT. Birgitta Törnkvist, Widar Henriksson

EM No 47. REPEATED TEST TAKING. Differences between social groups. Birgitta Törnkvist, Widar Henriksson

EM No 49. THE SWEDISH SCHOLASTIC ASSESSMENT TEST (SweSAT). Development, Results and Experiences. Christina Stage, Gunilla Ögren

EM No 50. CLASSICAL TEST THEORY VS. ITEM RESPONSE THEORY. An evaluation of the theory test in the Swedish driving-license test. Marie Wiberg

EM No 51. ENTRANCE TO HIGHER EDUCATION IN SWEDEN. Christina Stage

Em No 52. NOTES FROM THE TENTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 1–3, 2004. Christina Stage

2005

Em No 53. VALIDATION OF THE SWEDISH UNIVERSITY ENTRANCE SYSTEM. Selected results from the VALUTA-project 2001–2004. Kent Löfgren

Em No 54. SELF-ASSESSMENT OF KNOWLEDGE AND ABILITIES. A Litterature Study. Anna Sundström

2006

Em No 55. BELIEFS ABOUT PERCEIVED COMPETENCE. A literature review. Anna Sundström

Em No 56. VALIDITY ISSUES CONCERNING REPEATED TEST TAKING OF THE SWESAT. Birgitta Törnkvinst, Widar Henriksson

Em No 57. ECTS AND ASSESSMENT IN HIGHER EDUCATION. Conference Proceedings. Kent Löfgren

Em No 58. NOTES FROM THE ELEVENTH INTERNATIONAL SweSAT CONFERENCE. Umeå, June 12–14, 2006. Christina Stage

2007

Em No 59. PROCEEDINGS FROM THE CONFERENCE: THE GDE-MODEL AS A GUIDE IN DRIVER TRAINING AND TESTING. Umeå, May 7–8, 2007. Widar Henriksson, Tova Stenlund, Anna Sundström, Marie Wiberg

Em No 60. MEASURING AND DETECTING DIFFERENTIAL ITEM FUNCTIONING IN CRITERION-REFERENCED LICENSING TEST. A theoretic comparison of methods. Marie Wiberg