Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 6, April 2011 ISSN 1531-7714

Evaluating the Quantity-Quality Trade-off in the Selection of Anchor Items: a Vertical Scaling Approach

Florian Pibal and Hermann S. Cesnik
Alpen-Adria Universität Klagenfurt, Austria

When administering tests across grades, vertical scaling is often employed to place scores from different tests on a common overall scale so that test-takers' progress can be tracked. In order to be able to link the results across grades, however, common items are needed that are included in both test forms. In the literature there seems to be no clear agreement about the ideal number of common items. In line with some scholars, we argue that a greater number of anchor items bear a higher risk of unwanted effects like displacement, item drift, or undesired fit statistics and that having fewer psychometrically well-functioning anchor items can sometimes be more desirable. In order to demonstrate this, a study was conducted that included the administration of a reading-comprehension test to 1,350 test-takers across grades 6 to 8. In employing a step-by-step approach, we found that the paradox of high item drift in test administrations across grades can be mitigated and eventually even be eliminated. At the same time, a positive side effect was an increase in the explanatory power of the empirical data. Moreover, it was found that scaling adjustment can be used to evaluate the effectiveness of a vertical scaling approach and, in certain cases, can lead to more accurate results than the use of calibrated anchor items.

In order to ensure validity and fairness in scoring, the equating and linking of test scores is crucial for any testing program that produces new test forms and claims to deliver scores that have the same meaning over time. According to Holland (2006), a *link*, in general, is a transformation "between the scores from one test and those of another" (p. 5). Linking, or transforming scores, can be done in one of three ways: *predicting*, *scale aligning*, and *equating*. Typically, predicting involves the prediction of the score on one test (Y) on the basis of the score of another test (X). Scale aligning, or *scaling*, aims at placing the scores of two tests on a common scale, whereas equating represents a special case of scaling, in that a direct link is created between two test scores, resulting in test scores that are interchangeable. For the purpose of this paper, scale aligning (or scaling) will be of particular interest. Which of the approaches is employed in order to place scores on a common scale depends on several factors.

The focus of this article will be to evaluate the effects of a scaling approach that is widely used to link different test forms, namely *vertical scaling*, which usually involves tests of the same subject that are administered to different grades and aims to place scores on a common overall scale so that test-takers' progress can be tracked. Since there is sometimes great confusion in the use of linking terminology, for the purpose of this paper, *vertical scaling* will refer to a fixed-item parameter anchoring where adjacent grades have common items. The procedure to evaluate the vertical scaling will be referred to as *scaling adjustment* (cf. Cohen et al. 1993) and will involve the calculation of mean grade-to-grade differences of the adjacent grades in standard deviation units.

Although the findings of this study are applicable to all areas of testing, the specific subject of this paper is English materials used for the informal assessment of Austrian pupils at grades 6 and 7 in the four skills of reading, listening, writing and speaking. These tests are made publicly available to

teachers and used for diagnostic purposes. Designed as self-evaluation tools, this Diagnostic Profiling System General (DPSG) helps teachers prepare their students for the nationwide English Standards Tests (E8)[1] by providing an accurate diagnosis of where their students are in comparison to what is expected of them in the standards tests. Vertical scaling serves as an appropriate linking procedure since the DPSG and E8 tests are of the same length, thus exhibiting similar reliability, and comprise similar constructs, the main difference being overall test difficulty in order to match the lower population ability in the DPSG. In addition, from a content point of view, SLA acquisition orders are taken into account during the item-writing process. Typically, this multi-stage item-generation-and-feedback process includes (1) item writing, (2) multiple peer review, (3) multiple feedback, (4) item modification, and (5) final approval.

Before operational use, all the items for the diagnostic instruments were tested in small-scale pilot tests in order to study their psychometric properties. For this purpose, the Rasch Model was applied, as provided in Bond & Fox (2007: 45):

$$P_{ni} = \frac{e^{B_n - D_i}}{1 + e^{B_n - D_i}} \quad (1)$$

where B_n is the ability of person n and D_i is the difficulty of item i.

Thus, in the case of a dichotomous attainment item, P_{ni} is the probability of success upon interaction between the relevant person and the assessment item. Consequently, the log odds, or *logit*, of a correct response by a person to an item, is equal to $B_n - D_i$. The two most important properties of the Rasch Model are *invariant comparison* and *sufficiency*. The former means that the performance of two test takers should be comparable, regardless of which items they have solved while the latter means that all the information needed for such a comparison is contained in the person total (raw) score (Rasch, 1961: 332). If, however, tests are administered across grades, we need to put the person ability and item difficulty measures onto the same scale in order to guarantee that the requirement of invariant comparison is met.

Research Focus

Basic Considerations

In order to ensure invariant comparison in the case of test administration across grades, a number of linking procedures are available. One typical method is vertical scaling, i.e. the linking of test forms at adjacent grades through the use of common items (anchors) with fixed IRT item parameters. Many researchers insist that about 20-25% of the entire test length should be anchor items to get a better estimate of the common scale (Wright & Stone, 1979, Hambleton et al., 1991). Other scholars, like Smith and Kramer (1992), however, argue that even a single item can be employed as an anchor to place the scores from two different test forms onto a common scale. Before going into any depth about the test design and research methodology employed, some basic considerations about vertical scaling should be mentioned.

When anchor items are drawn from an existing bank and used for administration at a different grade than originally intended, some unforeseen complications might occur, including displacement, item drift, undesired fit statistics, or negative side effects like a decrease in the explanatory power of the model. If, for example, we use a high number of anchors and they all exhibit drift in the same direction, this will distort the measurement of the items contained in the new administration. Ideally, if we have a number of anchor items, unwanted effects like item drift or outlier-sensitive observations will usually average out over the total number of anchors. So, many anchors often merely serve the function of compensating for the weaknesses of the other anchor items.

In a case where many anchors are used, therefore, uncontrollable item drift might occur in one direction or another. In that case the variability – but not necessarily the amount of *variance explained* by the model, however – might increase and so the results of the second administration will not be accurate and consequently a weak basis for inference. Along the lines of Smith & Kramer (1992), we therefore argue that having fewer, well-functioning anchor items (in terms of their psychometric properties) can sometimes be more desirable than having a high number of anchors and we furthermore claim that, under certain conditions, scale alignment without calibration can lead to more accurate results than when using anchors.

Test Design

The reading-comprehension test comprised eight test forms, varying in difficulty and containing 20 multiple-choice items each. In each test form, there were four item blocks, consisting of five dichotomous items each. In test form 2, five items were originally intended as anchors in order to link the test results to the E8 item bank. Additionally, five shared items were used across all test forms in order to make the results of each test form comparable with each other[2] (cf. Table 1). In total, 1,350 persons and 115 items were tested.

[1] The English Standards Tests in Austria, taking place at grade 8, started in 2006, with baseline studies following in 2007, 2008, and 2009, and will take place every three years in the future.

[2] In terms of test design, the E8 test is similar to the DPSG design, differing only in the total number of test forms. Moreover, the anchor items in both tests are placed in the first item block. As a result, potential position effects can be ignored for the purpose of this paper.

Aim of the Study

This study aims to demonstrate that while a higher number of anchors might help obtain a better estimate of the common scale, it is not always the quantity but rather the quality of the anchors that is decisive in tying different scales together. Moreover, in certain cases, scaling adjustment, can lead to more accurate results than the use of anchors. The main reason for focusing on a reading-comprehension test is that the above hypothesis arose from the data collected in the context of the DPSG reading test, subsequently inspiring the following step-by-step approach.

As a first step, vertical scaling was employed by taking anchor items for the reading section from an existing item bank, i.e. a pool of already calibrated E8 items, in order to make the DPSG results projectable onto the E8 scale. After an initial analysis with five calibrated anchors, it turned out that they were not functioning satisfactorily. More precisely, two of them showed extraordinarily high displacement values. After eliminating the two problematic anchors, another analysis was conducted with the remaining three anchors, which led to an improvement in measurement quality. Finally, the vertical scaling approach was evaluated by calculating the separation of grade-to-grade distributions.

- Simulation 1 (vertical scaling using five anchors) was conducted to project the DPSG item difficulties onto the E8 scale using fixed-item parameter anchoring.
- Simulation 2 (vertical scaling using three anchors) aimed to produce more accurate parameter estimations for the DPSG item difficulties since twotems originally intended as anchors exhibited unsatisfactory item drift.
- Simulation 3 (scaling adjustment) was used to evaluate the results of vertical scaling on an IRT basis, demonstrating that scaling adjustment might produce more accurate results. For this purpose the effect size between the E8 and DPSG administrations was calculated in standard deviation units.

In the following, these three simulations with raw data from the reading pilot test will be discussed in terms of their effects on the person ability and item difficulty measures. The simulations were performed using the Rasch-modeling software WINSTEPS. Note that in terms of mean-square fit values, *all* estimations were in the range of 0.5 to 1.5, which can be considered "productive for measurement" (Wright & Linacre, 1994: 370) for the purpose of informal simulations.

Table 1: Design of the DPSG Reading-Comprehension Test

Test form 1	Test form 2	Test form 3	Test form 4-7	Test form 8
Standard_item_001	**Anchor_item_001**	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_002	**Anchor_item_002**	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_003	**Anchor_item_003**	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_004	**Anchor_item_004**	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_005	**Anchor_item_005**	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_006	*Standard_item_006*	*Standard_item_006*	*Standard_item_006*	*Standard_item_006*
Standard_item_007	*Standard_item_007*	*Standard_item_007*	*Standard_item_007*	*Standard_item_007*
Standard_item_008	*Standard_item_008*	*Standard_item_008*	*Standard_item_008*	*Standard_item_008*
Standard_item_009	*Standard_item_009*	*Standard_item_009*	*Standard_item_009*	*Standard_item_009*
Standard_item_010	*Standard_item_010*	*Standard_item_010*	*Standard_item_010*	*Standard_item_010*
Standard_item_011	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_012	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_013	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_014	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_015	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_***
Standard_item_016	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_106
Standard_item_017	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_107
Standard_item_018	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_108
Standard_item_019	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_109
Standard_item_020	Standard_item_***	Standard_item_***	Standard_item_***	Standard_item_110

Note that the items have been assigned generic names in order to facilitate identification throughout the paper.

Simulation 1: Vertical Scaling Using Five Anchors

Simulation 1 was performed with the raw data from the reading-comprehension test, with item entries 21-25 being defined as anchor items, and their item-bank values based on previous runs. In total, measures were generated for all 1,350 persons and 115 items. Additionally, the displacement values[3] for the anchor items were determined (see Table 2).

At first glance, the use of anchors generally results in higher item difficulty measures. On taking a closer look, we see that anchor items 001 and 002 show exceptionally high

[3] *Displacement values* indicate the difference between the observed (empirical) and the expected (anchor) score, i.e. how different the measures would be if they were not anchored.

discrepancies from their anchor values, i.e. their values obtained in previous administrations (cf. both their fit statistics and their displacement values). Large displacement values usually indicate that the item(s) affected should be unanchored. Consequently anchor values should be validated before they are used. For this purpose, two analyses were performed:

1. item and person measures were produced with no items anchored (i.e. all items *floating*), and
2. item and person measures were produced with the five provisional anchor items anchored (i.e. vertical scaling).

Next, a cross-plot was created for both the item difficulties and the person measures of the two runs. The cross-plotted item and person measures are shown in Figure 1.

As one would expect, the person measures form an almost straight line, indicating that a large number of respondents was not affected in their person measures. As for the item difficulties, the unanchored items form a straight line whereas some anchored items are noticeably off the line. Wright & Stone (1979) recommend that these are candidates for dropping as anchors. The effect of

Table 2: Item Difficulty Measures Using Five Anchors (Simulation 1)

Entry	Item	Measure	INFIT MNSQ	OUTFIT MNSQ	S.E.	PT-MSR CORR.	Displace-ment
1	Standard_item_001	-1.55	1.10	1.46	0.21	0.28	-
2	Standard_item_002	-1.10	0.96	0.89	0.14	0.47	-
3	Standard_item_003	-0.88	0.95	0.88	0.19	0.48	-
…	…	…	…	…	…	…	…
21	Anchor_item_001	-2.23	2.00	2.26	0.26	0.23	0.94
22	Anchor_item_002	2.11	1.79	2.31	0.23	0.17	-0.79
23	Anchor_item_003	-0.91	0.97	0.85	0.18	0.42	0.08
24	Anchor_item_004	-0.27	1.08	1.07	0.17	0.30	0.07
25	Anchor_item_005	0.57	1.17	1.26	0.17	0.25	-0.38
…	…	…	…	…	…	…	…
114	Standard_item_109	1.50	1.16	1.46	0.18	0.15	-
115	Standard_item_110	-0.80	0.94	0.87	0.17	0.45	-
	Mean	**0.08**	**0.99**	**1.01**	**0.18**		
	S.D.	**1.17**	**0.17**	**0.31**	**0.04**		

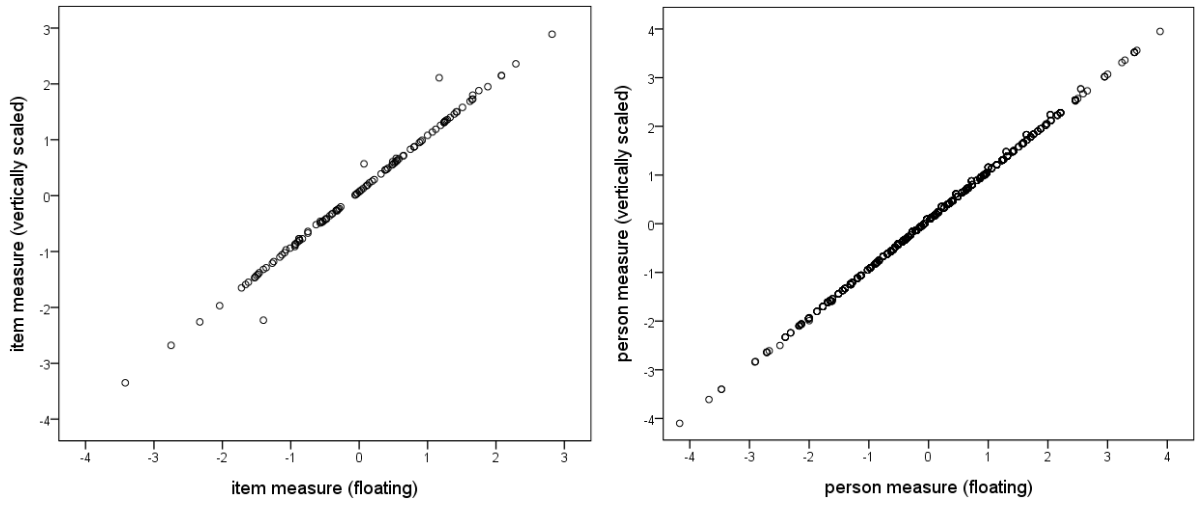

Figure 1: Item and Person Measure Cross-Plots: Vertical Scaling (Five Anchors) vs. Floating

unanchoring a displaced anchor item is to realign the person measures by roughly (displacement / number of remaining

anchored items). It is further suggested that random displacements of less than 0.5 logits are unlikely to have much impact in a test instrument (Wright & Stone, 1979: 98).

In addition to the cross-plot, a DIF analysis which was conducted to determine whether items would function differently according to the grade level they were administered at showed no significant differences in item difficulty between grades 6 and 7. However, the following paradoxical situation arose: across the items originally intended for eighth-grade students, the 'easier' items became more difficult for the sixth- and seventh-graders, and the 'more difficult' items became easier. This is a very common situation in vertical equating: Items that are too easy or too hard for a cohort tend to "drift". The item characteristic curves (ICCs) for the two problematic items are shown in Figure 2.

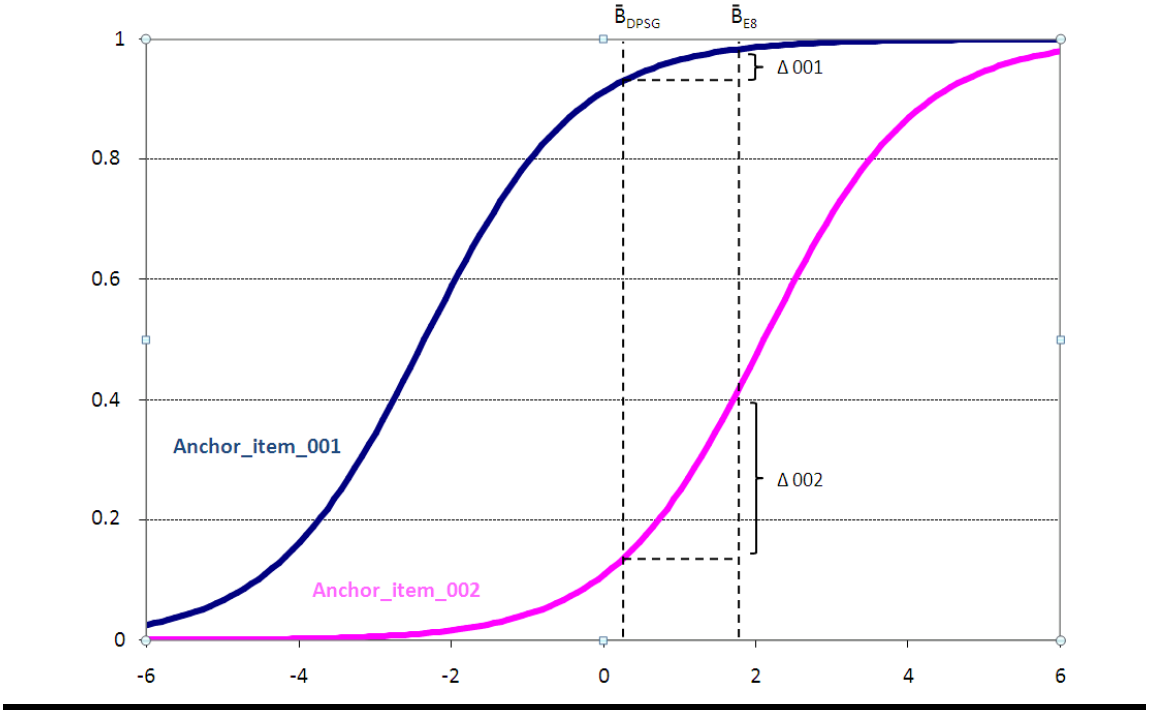

Figure 2: ICC for Anchor Items 001 and 002

The figure shows that the mean ability level for DPSG ($\bar{B}_{DPSG}$) lies at 0.16 logits whereas the mean for E8 ($\bar{B}_{E8}$) amounts to 1.96 logits. The two items are both on the flat part of the ICC, i.e. they do not discriminate well between the two different grades (different levels of person ability). Anchor_item_001 is 'too easy' resulting in a *ceiling effect* (Δ 001), whereas Anchor_item_002 is 'too difficult' resulting in a *threshold effect* (Δ 002). Furthermore, these two items exhibited the highest standard error (S.E.) of all anchor items (see Table 2).

Subsequently, a rank-order table of person ability measures for all 1,350 test takers was generated (see Table 3).

The person measures increase slightly in comparison to the results of the unanchored analysis, the reason for this being the different difficulty values and, hence, also the different relative positions of the five anchor items.

Simulation 1 shows that, in comparison to no anchor items being used, the item and person means shift up slightly by 0.08 logits, so that both the item difficulty and person ability measures will increase for all items and persons.

In order to identify potentially dependent items, standardized residual correlations were calculated for all items used. Since the items used as anchors were all short items which were each based on a different input text, the analysis showed, as expected, that there was no significant correlation among the anchor items, and thus no local dependence in the single testlets. Furthermore, no correlation values above 0.40 were detected for any of the other items, suggesting that each of the items used contributes meaningfully to measurement.

As for anchoring across grades, in many contexts the general rule is that items should not be administered more than one year away from their intended educational level, and even those items must be carefully selected (Ingebo, 1976). Consequently, on the basis of the results of Simulation 1, a vertical scaling approach using only three anchor items was implemented.

Table 3: Rank Order of Person Ability Measures Using Five Anchors (Simulation 1)

Rank	Entry	Person ID	Measure	INFIT MNSQ	OUTFIT MNSQ	S.E.	PT-MSR CORR.
1	355	Person_0355	3.95	1.02	0.69	1.04	0.17
2	1124	Person_1124	3.56	0.94	0.58	1.04	0.31
3	643	Person_0643	3.52	0.81	0.69	1.05	0.36
…	…	…	…	…	…	…	…
662	687	Person_0687	0.17	0.79	0.73	0.50	0.61
663	699	Person_0699	0.17	1.24	1.25	0.50	0.27
…	…	…	…	…	…	…	…
1349	685	Person_0685	-4.86	1.43	1.49	1.88	0.00
1350	688	Person_0688	-4.86	1.43	1.49	1.88	0.00
		Mean	**0.16**	**1.00**	**1.04**	**0.55**	
		S.D.	**1.15**	**0.21**	**0.45**	**0.12**	

Simulation 2: Vertical Scaling Using Three Anchors

Since in the previous simulation, two anchor items showed exceptionally high displacement values, another simulation was conducted, eliminating these two problematic items as anchors. The item statistics are shown in Table 4 and the person statistics are shown in Table 5.

Here both, item and person measures shift up slightly, resulting in a new mean for both. As already mentioned, the effect of unanchoring the displaced anchor items is to realign the person measures by roughly (displacement / number of remaining anchored items), thus producing more reliable and generalizable parameter estimates for the DPSG item difficulties. Whereas in some contexts anchoring with only three anchor items can be problematic, here an improvement in measurement quality was achieved by *reducing* the number of anchors. Consequently, an evaluation procedure was carried out to assess the results of vertical scaling on an IRT basis and – to go one step further – demonstrate that, in the case of bad-fitting anchor items, scaling adjustment can produce more accurate results.

Table 4: Item Difficulty Measures Using Three Anchors (Simulation 2)

Entry	Item	Measure	INFIT MNSQ	OUTFIT MNSQ	S.E.	PT-MSR CORR.	Displace-ment
1	Standard_item_001	-1.51	1.10	1.45	0.21	0.28	-
2	Standard_item_002	-1.06	0.96	0.89	0.14	0.47	-
3	Standard_item_003	-0.84	0.95	0.88	0.19	0.48	-
…	…	…	…	…	…	…	…
21	*former* Anchor_item_001	-1.26	1.07	1.07	0.20	0.23	-
22	*former* Anchor_item_002	1.31	1.15	1.25	0.19	0.17	-
23	Anchor_item_003	-0.91	0.98	0.86	0.18	0.42	0.08
24	Anchor_item_004	-0.27	1.08	1.06	0.17	0.30	0.07
25	Anchor_item_005	0.57	1.15	1.23	0.17	0.25	-0.38
…	…	…	…	…	…	…	…
114	Standard_item_109	1.54	1.16	1.46	0.18	0.15	-
115	Standard_item_110	-0.76	0.94	0.87	0.17	0.45	-
	Mean	**0.12**	**0.97**	**0.99**	**0.18**		
	S.D.	**1.14**	**0.13**	**0.26**	**0.04**		

Table 5: Rank Order of Person Ability Measures Using Three Anchors (Simulation 2)

Rank	Entry	Person ID	Measure	INFIT MNSQ	OUTFIT MNSQ	S.E.	PT-MSR CORR.
1	355	Person_0355	3.99	1.02	0.69	1.04	0.17
2	1124	Person_1124	3.60	0.94	0.58	1.04	0.31
3	643	Person_0643	3.56	0.81	0.69	1.05	0.36
…	…	…	…	…	…	…	…
675	687	Person_0687	0.21	0.79	0.73	0.50	0.61
676	699	Person_0699	0.21	1.24	1.25	0.50	0.27
…	…	…	…	…	…	…	…
1349	685	Person_0685	-4.82	1.43	1.49	1.88	0.00
1350	688	Person_0688	-4.82	1.43	1.49	1.88	0.00
		Mean	**0.20**	**1.00**	**1.03**	**0.55**	
		S.D.	**1.15**	**0.21**	**0.42**	**0.12**	

Simulation 3: Scaling Adjustment

As a final step, Kolen & Brennan (2004) mention three ways of evaluating the vertical scale that has been created in the previous simulation: *grade-to-grade growth* (i.e. differences in the means or medians of score distributions at adjacent grades), *grade-to-grade variability* (i.e. differences in the standard deviations or other variability measures), and *separation of grade distributions*. The first two can be performed by visual inspection of the growth curve plots, whereas the separation of grade distributions is determined by calculating the effect size (i.e. the mean grade-to-grade differences in standard deviation units) of a vertical scaling approach. Since it takes both means and variability into account, the last approach will be the focus of this section. It consists of three stages:

1. calculating the effect size in the difficulty of the anchor items between grades 6/7 (DPSG) and grade 8 (E8) in standard deviation units
2. converting the effect size into logits
3. adjusting the DPSG scale accordingly in order to project the results of the DPSG test onto the E8 scale (which is also the common scale in the item bank)

As already seen in Simulations 1 and 2, the item difficulty measures increased slightly from E8 to DPSG, which means that both the item difficulty and the person ability mean are slightly higher than without anchoring. The situation is illustrated in Figure 3.

Calculation of effect size in standard deviation units

First, the effect size of the grade-to-grade differences was calculated for each of the five items available in both administrations. For this purpose, Kolen & Brennan (2004: 410) suggest using an index first proposed by Yen (1986):

$$eff.size\ (S.D.) = \frac{\bar{x}_{upper} - \bar{x}_{lower}}{\sqrt{\frac{n_{upper} \cdot s^2_{upper} + n_{lower} \cdot s^2_{lower}}{n_{upper} + n_{lower}}}} \quad (2)$$

$$eff.size\ (S.D.) = \frac{\bar{x}_{E8} - \bar{x}_{DPSG}}{\sqrt{\frac{n_{E8} \cdot s^2_{E8} + n_{DPSG} \cdot s^2_{DPSG}}{n_{E8} + n_{DPSG}}}} \quad (3)$$

Applying Equations (2) and (3) to DPSG and E8 gave the results shown in Table 6.

Note that the effect size is a weighted value since the sample size for each item is different. However, it is already obvious that an effect size of 0.049 standard deviation units will not shift the item difficulty mean too greatly. Nonetheless, such marginal differences in item difficulty could still affect not only the item statistics in the item bank, but also students' placement on the scale, which might in turn influence their classification according to the cut scores. Thus, it would be more difficult to track item behavior across administrations as well as students' progress across grades.

Conversion of effect size into logits

As a next step, the effect sizes for each item and the average effect size were converted into logits, as illustrated in Equation (4):

$$eff.size(LOG) = eff.size(S.D.) \cdot S.D.(persssample(LOG)) \quad (4)$$

The result is shown in Table 7.

Adjustment of the DPSG scale

Finally, the mean difficulty measure of the DPSG scale was adjusted according to the average weighted effect size in logits. A WINSTEPS simulation was performed and a new

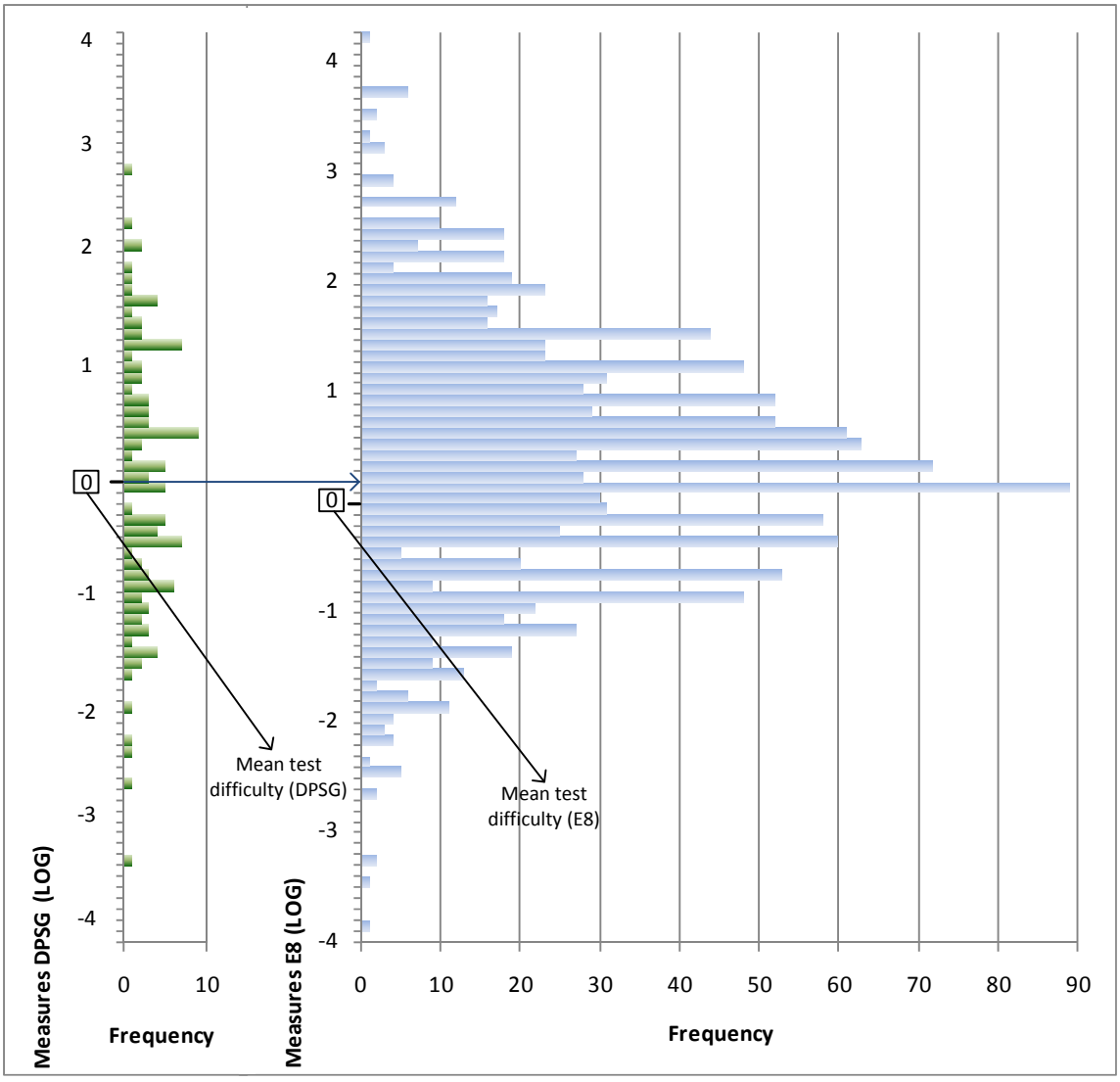

Figure 3: Projection of DPSG Item Difficulties onto the E8 Scale Using Scaling Adjustment

Table 6: Weighted Effect Size for Common Items in Standard Deviation Units

Item	Item measure (LOG)	$\bar{x}_{E8}$	$\bar{x}_{DPSG}$	$\Delta(\bar{x}_{E8} - \bar{x}_{DPSG})$	$)s^2_{E8}$	s^2_{DPSG}	n_{E8}	n_{DPSG}	effect size (S.D. units)
Anchor_item_001	-2.23	0.80	0.78	0.02	0.161	0.173	517	168	0.00756949
Anchor_item_002	2.11	0.31	0.27	0.04	0.215	0.200	719	168	0.01723239
Anchor_item_003	-0.91	0.84	0.70	0.14	0.132	0.210	225	168	0.03027049
Anchor_item_004	-0.27	0.77	0.58	0.19	0.178	0.245	209	168	0.03514838
Anchor_item_005	0.57	0.65	0.49	0.16	0.226	0.251	1960	168	0.15952976
Mean effect size (weighted – S.D. units):									**0.04995010**

Table 7: Weighted Effect Size for Common Items in Logits

Items	effect size (S.D. units)	S.D. of person sample (LOG)	Effect size (LOG)
Anchor_item_001	0.00756949	1.20	0.00908338
Anchor_item_002	0.01723239	1.20	0.02067886
Anchor_item_003	0.03027049	1.20	0.03632459
Anchor_item_004	0.03514838	1.20	0.04217805
Anchor_item_005	0.15952976	1.20	0.19143571
Mean effect size (LOG):			**0.05994012**

Table 8: Item Difficulty Measures Using Scaling Adjustment (Simulation 3)

Entry	Item	Measure	INFIT MNSQ	OUTFIT MNSQ	S.E.	PT-MSR CORR.
1	Standard_item_001	-1.56	1.10	1.45	0.21	0.28
2	Standard_item_002	-1.10	0.96	0.89	0.14	0.47
3	Standard_item_003	-0.88	0.95	0.88	0.19	0.48
…	…	…	…	…	…	…
21	Anchor_item_001	-1.34	1.07	1.07	0.20	0.23
22	Anchor_item_002	1.23	1.15	1.25	0.19	0.17
23	Anchor_item_003	-0.88	0.94	0.82	0.18	0.42
24	Anchor_item_004	-0.26	1.06	1.04	0.17	0.30
25	Anchor_item_005	0.13	1.09	1.14	0.17	0.25
…	…	…	…	…	…	…
114	Standard_item_109	1.48	1.17	1.46	0.18	0.15
115	Standard_item_110	-0.81	0.94	0.87	0.17	0.45
	Mean	**0.06**	**0.97**	**0.99**	**0.18**	
	S.D.	**1.14**	**0.13**	**0.26**	**0.04**	

Table 9: Rank Order of Person Ability Measures Using Scaling Adjustment (Simulation 3)

Rank	Entry	Person ID	Measure	INFIT MNSQ	OUTFIT MNSQ	S.E.	PT-MSR CORR.
1	355	Person_0355	3.94	1.02	0.69	1.04	0.17
2	1124	Person_1124	3.55	0.95	0.58	1.04	0.31
3	643	Person_0643	3.51	0.81	0.69	1.05	0.36
…	…	…	…	…	…	…	…
675	687	Person_0687	0.16	0.79	0.73	0.50	0.61
676	699	Person_0699	0.16	1.24	1.25	0.50	0.27
…	…	…	…	…	…	…	…
1349	685	Person_0685	-4.87	1.43	1.49	1.88	0.00
1350	688	Person_0688	-4.87	1.43	1.49	1.88	0.00
		Mean	**0.14**	**1.00**	**1.02**	**0.54**	
		S.D.	**1.14**	**0.21**	**0.42**	**0.12**	

Table 10: Comparison of the Three Simulations and their Impact on Means & Variability

	Vertical scaling (5 anchors)	Vertical scaling (3 anchors)	Scaling adjustment
Mean (S.D.)	0.08 (1.17)	0.12 (1.14)	0.06 (1.14)
INFIT Mean-square (S.D.)	0.99 (0.17)	0.97 (0.13)	0.97 (0.13)
OUTFIT Mean-square (S.D.)	1.01 (0.31)	0.99 (0.26)	0.99 (0.26)
Variance explained by measures (modeled)	48.7 (48.6)	49.8 (49.3)	49.8 (49.3)
Variance explained by persons (modeled)	24.7 (24.6)	24.9 (24.6)	24.9 (24.6)
Variance explained by items (modeled)	24.0 (24.0)	24.9 (24.7)	24.9 (24.7)

item mean of 0.06 logits was specified so that all measures were increased by 0.06 logits. Again, measures for all 1,350 persons (see Table 8) and 115 items (see Table 9) were generated.

Discussion

In comparison to Simulation 2, Simulation 3 shows almost identical person measures, with a maximum deviation of 0.01 logits. The same goes for most items – with the exception of the five anchor items which, unlike in Simulation 1, do not exhibit high displacement values. If we compare the three approaches in terms of their impact on means and variability, we get the picture shown in Table 10.

We can see that both eliminating the two problematic anchors as well as shifting the mean by the effect-size correction factor have several effects. First, both procedures lead to a slight decrease in standard deviation. Second, the fit statistics shift to some extent, leading to a slight model overfit. Moreover, the two approaches result in reduced variability in terms of OUTFIT statistics, which are most sensitive to outlying observations. This can be explained by the use of the two problematic items as anchors. Notably, we can also see that by unanchoring the two items with high item drift, we achieve a slight gain in the explanatory power of the model. The total raw variance explained by the measures amounts to 48.7%, which exceeds the model by 0.1%. In case of Simulations 2 and 3, the degree of variance explained in the observations is 49.8%, which is higher than in Simulation 1, and exceeds the modeled amount by 0.5%. Finally, we can see that the percentage of variance explained (either modeled or empirical), is always higher when the two unstable anchors are dropped. With the main intention being to achieve a better referencing of the DPSG test to the E8 test, it is important to note that the positive side effect of an increase in the amount of variance explained depends, in turn, on the goodness and variance of the excluded anchor items. Thus, since there was only a slight increase in variance explained, we conclude that the former anchors exhibited low variance.

A subsequent dimensionality analysis further indicated that there was no clear secondary dimension present in the data. Together with the fact that no excessive amount of misfitting items or persons was detected, this leads to the assumption that the data are under statistical control and that the amount of "variance explained" is satisfactory given the sample and the instrument. To sum up, *not* using the five original anchor items leads to more accurate measures because the vertical scaling paradox of extremely easy/difficult items "drifting" towards the middle is avoided.

Summary and Recommendations

This study aimed to demonstrate that it is not always the quantity but rather the quality of the anchors that is decisive in tying different scales together. By employing a step-by-step approach, we found that the paradox of high item drift in test administrations across grades can be mitigated, and possibly even be eliminated while at the same time, increasing the explanatory power of the empirical data. Moreover we found that scaling adjustment can be used to evaluate the effectiveness of a vertical scaling approach and, in certain cases, can lead to more accurate results than the use of calibrated anchor items.

Based on the results of our study and in line with previous research (Smith & Kramer, 1992), we claim that – under certain premises – in the case of vertical scaling of test results across different grades, having fewer, psychometrically well-functioning anchor items can be more desirable than having a larger number of unstable anchors. In other words, if stable anchor items are available, the application of fewer anchors is sufficient (and in the most extreme case, hypothetically, one such stable anchor would suffice). For researchers and practitioners who need to compare assessments across grades, this means when anchoring across grades, as recommended in the literature, it is important in many contexts that items should not be administered more than one year away from their intended educational level and that even those items must be carefully selected (Ingebo, 1976). One strong indicator for the goodness of anchor items may be the item measure (provided that the chosen items show acceptable fit statistics, of course). It is desirable not to use items with extremely high or extremely low measures since they are more susceptible to drift, which can result in undesirably high displacement values. In practice this means that if an initial analysis yields problematic results, it is worth considering a reduction in the number of anchor items. As demonstrated, eliminating problematic anchors led to an improvement in measurement quality. Given the quality of

the three remaining anchors, it is important to note that the low number of items suffices to justify making appropriate inferences. Still, such a low number merely serves for demonstration purposes whereas, in reality, a higher number of anchor items will be more desirable, especially for high-stakes tests. Should it be the case that none of the intended anchors is functioning well, scaling adjustment, i.e. calculating the separation of grade-to-grade distributions, can be applied to project the results of different test administrations across grades onto the same scale. Above and beyond that, the scaling adjustment approach might be used to evaluate the results of a vertical scaling procedure.

However, we must also bear in mind objections that have been raised to vertical scaling. Schafer (2006), for example, lists some, pointing out, among other things, that "scores for students in lower grade levels are overestimated due to lack of data about inabilities over contents at higher grade levels" (p. 2), which could explain the paradox of (anchor) item drift across grades and consequently explain the "decrease in student scores from year-to-year" or "negative growth" (p. 3). Further limitations of vertically-scaled tests are mentioned by Kolen & Brennan (2004). They state that due to a lack of items in extreme scale score regions, "the psychometric comparability of scale scores (…) [across grades] (…) is limited to the range of scores at or below" (p. 412) the maximum score on either of the tests. Moreover, they point out that "content differences for the tests lead to limitations on the meaning of test scores" (ibid.), which means that in the case of dissimilar constructs, unidimensionality violations can occur and other methods such as *battery scaling* or *scaling on a hypothetical population* (Holland, 2006) might be more appropriate.

Finally, comparison of the item drift examined here to the drift found in replication across years as well as identification of influential factors that make items drift when making use of anchoring across grades might be potential avenues for further research. Moreover, future studies could include a comparison of the effects of vertical scaling (1) to simple mean equating based on raw scores, (2) to other IRT models (e.g. using Stocking-Lord adjustment), and, in this context also, (3) to *prediction* types of (IRT) linking. Additional research could also include an investigation of how to link the results of such vertical scaling approaches to an absolute framework of language competence such as the CEFR. For this purpose a variety of standard-setting approaches has been developed:

- traditional, CTT- or IRT-based, approaches like the Angoff or the Bookmark procedure (cf. Cizek & Bunch, 2007)
- linking methods that seek to add a level of comparability across different assessments (grades) like vertically-moderated standards (Lissitz & Huynh, 2003) or growth scales (Schafer & Twing, 2006)
- doubly IRT-based approaches, considering item calibrations based on test-taker performance and raters' estimates of item difficulty (Sigott & Cesnik, 2010).

References

Bond, T. and C. Fox (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. London: Lawrence Earlbaum Associates.

Cizek, G.J. and M.B. Bunch (2007). *Standard Setting. A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks: Sage Publications.

Cohen, A. S., Kim, S. H., and F.B. Baker (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement,* 17, 335–350.

Hambleton, R. K., Swaminathan, H., and J.H. Rogers (1991). *Fundamentals of item response theory*. New York: Sage Publications.

Holland, P. (2006). A Framework and History for Score Linking. Dorans, N.J., Pommerich, M. and P.W. Holland (eds.): *Linking and Aligning Scores and Scales*. New York: Springer.

Ingebo, G. (1976). How to link tests to form an item pool. Paper presented to *American Educational Research Association.* San Francisco, 1976.

Kolen, M.J. and R.L. Brennan (2004). *Test equating, scaling, and linking: methods and practices*. New York: Springer.

Linacre, John M. (2008). *Winsteps Rasch Measurement Computer Program.* Chicago: Winsteps.com.

Lissitz, R. W. and H. Huynh (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation,* 8(10).

Rasch, G. (1961). "On general laws and the meaning of measurement in psychology", in: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley: University of Chicago Press, 1980, 321-334.

Schafer, William D. (2006). Growth Scales as an Alternative to Vertical Scales. *Practical Assessment Research & Evaluation*, 11(4).

Sigott, G. and H. Cesnik (2010). Linking the Klagenfurt Item Bank to the CEFR. *Sprachlehrforschung: Theorie und Empirie – Festschrift für Rüdiger Grotjahn*. Frankfurt: Peter Lang, 147-156.

Smith, R.M. and G.A. Kramer (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement*, 52 (4), 835-846.

Wright, B.D. and M.H. Stone (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, B.D. and J.M. Linacre (1994). *Rasch Measurement Transactions*, 1994, 8:3, 370.

Citation:

Pibal, Florian & Hermann S. Cesnik (2011). Evaluating the Quantity-Quality Trade-off in the Selection of Anchor Items: a Vertical Scaling Approach. *Practical Assessment, Research & Evaluation*, 16(6). Available online: http://pareonline.net/getvn.asp?v=16&n=6.

Authors:

Florian Pibal
Language Testing Centre
Department of English and American Studies
Alpen-Adria Universität Klagenfurt
Universitätsstr. 64, Haus 13
A-9020 Klagenfurt, AUSTRIA

florian.pibal [at] uni-klu.ac.at
http://www.uni-klu.ac.at/ltc

Hermann S. Cesnik
Language Testing Centre
Department of IT Services
Alpen-Adria Universität Klagenfurt
Universitätsstr. 64, Haus 13
A-9020 Klagenfurt, AUSTRIA

hermann.cesnik [at] uni-klu.ac.at
http://www.uni-klu.ac.at/ltc