Supplement to: A framework for anchor methods and an iterative forward approach for DIF detection

Julia Kopf, Ludwig-Maximilians-Universität München
Achim Zeileis, Universität Innsbruck
Carolin Strobl, Universität Zürich

This supplement provides additional examples and explanations to accompany the main article *A framework for anchor methods and an iterative forward approach for DIF detection*. We first illustrate the requirement that the anchor items should be DIF-free by means of an instructive example, before we provide the background and motivation for our simulation study, as well as additional results of our simulation study by means of additional figures, tables and summaries not shown in the main article.

A. An instructive example

The data set from a general knowledge quiz was conducted by the weekly German news magazine SPIEGEL in 2009. A thorough discussion and analysis of the original data set are provided in Trepte and Verbeet (2010) including a global DIF analysis by means of model-based recursive partitioning by Strobl, Kopf, and Zeileis (2010). From about 700,000 test-takers that answered each a total of 45 items from different domains, we select a subsample of $9,442$ test-takers (that obtained their A-levels in Germany) and four items from politics (listed below together with the correct answers) for the illustration of the anchor problem:

Item 1	Who determines the rules of action in politics according to the German Constitution? (The Bundeskanzler.)
Item 2	What is the role of the second vote in the elections for the German Bundestag? (It governs the seating in the German Bundestag.)
Item 3	How many people were killed by the RAF? (33)
Item 4	Indicate the location of Hessen on the German map.

As an exemplary illustration, let us suppose we want to test for DIF in the first item between the focal (foc) group of the test-takers that obtained their A-levels in the German federal state Hessen and the reference (ref) group of all remaining test-takers. Figure A.1 displays three different restrictions: The second item as constant single-anchor, the fourth item as constant single-anchor and all other items (item 2 to item 4) as anchor. The points represent the estimated item parameters from the reference (light points) and the focal group (dark points). The rectangles surround the anchor item(s).

In Figure A.1 (left), item 2 is used as constant single-anchor and, thus, both estimated item parameters are set to zero. The negligible difference in the item parameters of item 1, that we are currently interested in, suggests no DIF in this item. The item-wise Wald test (see equation 5 in the main article) for item 1 does not display statistically significant DIF ($t = -.968$ with the corresponding p-value of .333). As a result, item 1 is classified as DIF-free. To understand the DIF test results for item 1 in the next scenarios, it is also important to note that the large difference in item 4 implies DIF in this item. Since item 4 was the question to indicate the location of Hessen on the German map, it is plausible that this item 4 is a true DIF item since it was easier for test-takers that obtained their A-levels in Hessen.

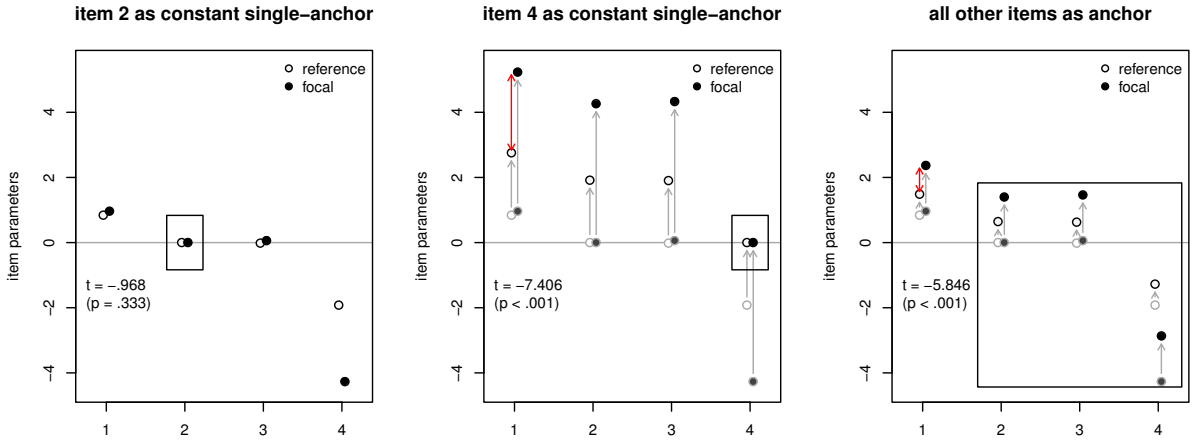

Figure A.1: Different restrictions placed on the item parameters that are estimated using the Rasch model in each group.

In the next scenario in Figure A.1 (middle), item 4 (that we just found plausible to have true DIF) is used as a constant single-anchor. Compared to the first scenario, all item parameters are now shifted upwards by the estimated difficulties of item 4 and artificial differences occur for item 1, 2 and 3. This shows that the anchor items should be DIF-free to avoid the artificial differences, which are termed artificial DIF by Andrich and Hagquist (2012). The artificial DIF for item 1, that we are currently interested in, is statistically significant ($t = -7.406$ with the corresponding p-value $< .001$). Hence, item 1 is classified as a DIF item.

In the last scenario in Figure A.1 (right), all other items – except the currently studied item 1 – are used as anchor items. Compared to the second scenario, the scales are shifted apart less strongly since the scale shift is reduced from the estimated difficulties of the DIF item 4 to the average over the estimated difficulties of item 2, 3 and 4 (including the apparently DIF-free items 2 and 3) as visible by the shorter arrows. However, the statistical test still classifies item 1 as a DIF item ($t = -5.846$ with the corresponding p-value $< .001$).

This example illustrates the major impact of the anchor method on the results of the DIF analysis, since – depending on the anchor set – three different test statistics result in the DIF tests for item 1.

B. Background and motivation of the simulation study

In this section, the background of our simulation study – that investigates the trade-off between the false alarm rate and the hit rate of DIF tests using the anchor methods introduced in Section 3 in our main article – is described. The results are used to develop guidelines which anchor methods should be used for DIF analysis in the Rasch model.

If no DIF is present in the test, we expect all anchor methods to yield well-controlled false alarm rates, since no DIF items and, therefore, no risk of contamination exists (Wang and Yeh 2003; Stark, Chernyshenko, and Drasgow 2006; Woods 2009; González-Betanzos and Abad 2012).

If DIF is balanced, i.e. the DIF items favor either the reference or the focal group and no systematic disadvantage exists, previous simulation studies showed that the all-other class yielded a well-controlled false alarm rate and a high hit rate (Wang and Yeh 2003; Wang 2004). However, if DIF is unbalanced i.e. all DIF items are simulated to favor one group, an inflated false alarm rate for the all-other method was reported (Wang and Yeh 2003; Wang 2004).

In accordance with Thissen, Steinberg, and Wainer (1988) and Woods (2009), we anticipate the constant anchor class to show an increase in the false alarm and the hit rate when the anchor length rises from one to four items and the proportion of DIF items is high. Wang, Shih, and Sun (2012) also found that four anchor items combined with the IRTLRDIF procedure (Thissen 2001) yielded low power rates as might also be the case in our simulation with the Wald test.

González-Betanzos and Abad (2012) compared an iterative backward two-step procedure based on the AO-selection strategy to specific constant single-anchors, to a purification procedure based on a DIF-free constant single anchor and to the all-other method. The constant single-anchor items were selected from the set of known a priori DIF-free items. The iterative backward two-step procedure showed slightly inflated false alarm rates. Due to the fact that one additional purification step improved the test results, the authors assumed improvements when further purification steps are added as we have implemented in our main article. Accordingly, we expect the iterative backward anchor class to achieve high hit rates as they allow for a long anchor, but at the expense of an inflated false alarm rate especially in settings where the proportion of DIF items is high and DIF is unbalanced.

Little information is available on how well the anchor selection strategies perform, as Wang

and Yeh (2003), Wang (2004) and Thissen *et al.* (1988) included only DIF-free items in the constant anchor class. This approach is only possible in simulation studies, however, where it is known by design which items are DIF-free. In practice, on the other hand, a set of DIF-free items prior to DIF analysis is usually not available (González-Betanzos and Abad 2012). Including only DIF-free items avoids the risk of contamination (for the consequences of contamination see Section *The anchor process for the Rasch model* in our main article and the empirical example in this supplement) and, thus, leads to an advantage for the methods from the constant anchor class. However, in order to compare the anchor classes under realistic conditions where it is not known a priori which items are DIF-free, the methods from the constant anchor class should be investigated together with an anchor selection strategy.

Woods (2009) investigated the AO-selection strategy to locate a set of constant anchor items and found results suitable for DIF analysis and superior to the all-other method. However, Wang *et al.* (2012) investigated the constant anchor method based on the selection of four anchor items using the AO-selection strategy (here referred to as the four-anchor-AO method) and found that the anchors were often contaminated and showed an inflated false alarm rate when DIF was unbalanced and no additional purification step was used. Therefore, we expect the four-anchor-AO method to perform well only in the condition of balanced DIF and poorly in the condition where DIF is unbalanced (Wang and Yeh 2003; Wang 2004; Shih and Wang 2009; Wang *et al.* 2012).

The SA-selection strategy proposed by Wang (2004) is (to our knowledge) implemented and combined with several anchor classes in our main article for the first time. Since the SA-selection strategy relies on DIF tests using every item as single anchor, we anticipate the SA-selection strategy to outperform the AO-selection strategy if the sample size is large and DIF is unbalanced. When DIF is balanced, we expect the AO-selection strategy to be superior.

The newly suggested iterative forward class builds the anchor in a step-by-step forward procedure. In comparison with the iterative backward method, we expect the forward procedure to be superior when the SA-selection strategy is used and DIF is unbalanced since the initial step of the iterative backward procedure is built on biased test results. In comparison with methods from the constant anchor class, we anticipate higher hit rates because the anchor of the iterative forward procedure grows as long as the current anchor is shorter than the number of currently presumed DIF-free items and should, thus, include more than four items. As a drawback, we also expect higher false alarm rates since the risk of contamination increases with the anchor length. Furthermore, we anticipate the methods from the iterative forward class to show lower hit rates than the all-other method in the balanced case, because the latter uses all items – except the studied item – as anchor.

C. Additional results of our simulation study

In this section, we provide additional results from our simulation study by means of additional figures, tables and summaries.

Null hypothesis: No DIF

Since all items were truly DIF-free in the first condition, only the false alarm rate (proportion of DIF-free items that were diagnosed with DIF) was computed.

False alarm rates

The estimated false alarm rates are depicted in Figure C.1 and (only for equal sample sizes) also reported together with their standard errors in Table C.1.

As shown in Figure C.1, all anchor methods held the 5% level. While methods from the all-other, the iterative backward (iterative-backward-AO) and the iterative forward class (iterative-forward-SA, iterative-forward-AO) together with the constant four-anchor-NC method were near the significance level of 5%, most methods from the constant anchor class (constant single-anchors: single-anchor-AO and single-anchor-SA; constant four-anchors: four-anchor-AO and four-anchor-SA) remained below that level. The constant single-anchors – that consist of an anchor with the constant length of only one item – displayed false alarm rates not exceeding 0.01, whereas the constant four-anchors displayed slightly higher false alarm rates (approximately 0.03 for the constant four-anchor-AO as well as for the four-anchor-SA method).

false alarm rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
no DIF									
250, 250	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.03)	(0.00)	(0.03)	(0.04)	(0.04)
500, 500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.03)	(0.01)	(0.02)	(0.03)	(0.03)	(0.00)	(0.03)	(0.04)	(0.03)
750, 750	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.03)	(0.01)	(0.03)	(0.03)	(0.03)	(0.00)	(0.03)	(0.04)	(0.04)
1000, 1000	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.03)	(0.01)	(0.02)	(0.03)	(0.03)	(0.00)	(0.03)	(0.04)	(0.03)
1250, 1250	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.00)	(0.03)	(0.04)	(0.04)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.03)	(0.00)	(0.03)	(0.04)	(0.04)

Table C.1: False alarm rates and standard errors under the null hypothesis (no DIF) with equal sample sizes in reference and focal group.

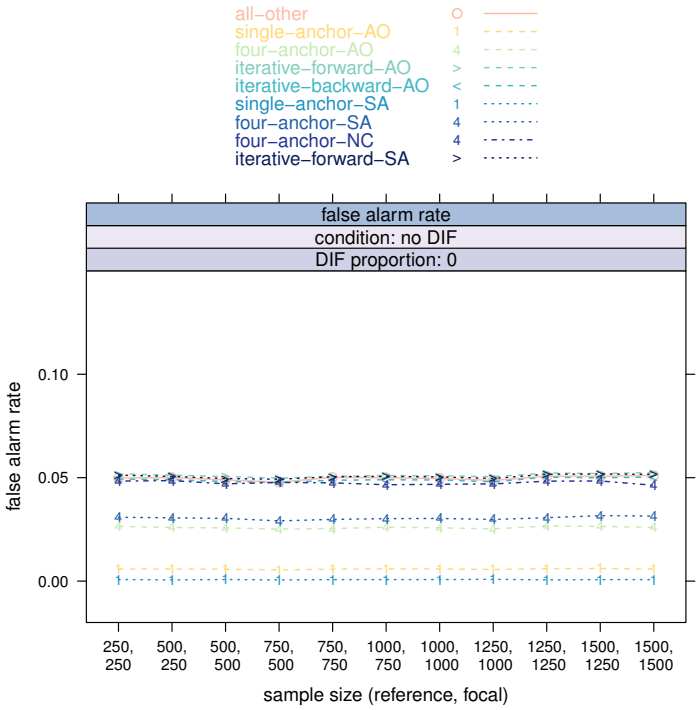

Figure C.1: False alarm rates under the null hypothesis of no DIF.

Hence, DIF tests with an anchor method from the constant anchor class combined with the AO- and the SA-selection – especially the constant single-anchor methods – were over-conservative.

Balanced DIF: No advantage for one group

The false alarm rates and hit rates for the balanced condition are presented in Figure 1 in our main article together with a detailed interpretation of the results. In addition, here, the false alarm rates are listed together with the standard errors only for equal sample sizes in Table C.2. The hit rates are included in Table C.3.

Summary

In the balanced condition, the AO-selection strategy outperformed the SA-selection by yielding higher hit rates as expected. The difference was large for methods from the constant anchor class, but negligible for methods from the iterative forward anchor class.

All anchor methods showed a well-controlled false alarm rate, except the constant four-anchor-SA and the four-anchor-NC method. All iterative methods (from the forward and backward class) and the all-other method displayed the most rapidly rising hit rates. The newly suggested iterative-forward-AO and iterative-forward-SA method enabled a high rate of correctly classified DIF items and simultaneously maintained the significance level in the balanced condition.

false alarm rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250, 250	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	(0.05)	(0.04)
500, 500	0.05	0.01	0.03	0.05	0.05	0.00	0.04	0.08	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.07)	(0.04)
750, 750	0.05	0.01	0.03	0.05	0.05	0.01	0.04	0.10	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.08)	(0.04)
1000, 1000	0.05	0.01	0.03	0.05	0.05	0.00	0.04	0.08	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.07)	(0.04)
1250, 1250	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.06	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	(0.06)	(0.04)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.06	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	(0.05)	(0.04)
0.30									
250, 250	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.02)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.06)	(0.04)
500, 500	0.05	0.01	0.03	0.05	0.05	0.01	0.05	0.11	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.05)	(0.10)	(0.04)
750, 750	0.05	0.01	0.03	0.05	0.05	0.02	0.06	0.14	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.11)	(0.04)
1000, 1000	0.05	0.01	0.03	0.05	0.05	0.01	0.04	0.12	0.05
	(0.04)	(0.02)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.10)	(0.04)
1250, 1250	0.05	0.01	0.03	0.05	0.05	0.01	0.04	0.08	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.04)	(0.08)	(0.04)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.07	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.07)	(0.04)
0.45									
250, 250	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.06	0.05
	(0.05)	(0.02)	(0.03)	(0.05)	(0.05)	(0.01)	(0.04)	(0.07)	(0.05)
500, 500	0.05	0.01	0.03	0.05	0.05	0.02	0.06	0.17	0.06
	(0.05)	(0.02)	(0.04)	(0.05)	(0.05)	(0.04)	(0.07)	(0.16)	(0.05)
750, 750	0.05	0.01	0.03	0.05	0.05	0.05	0.09	0.21	0.06
	(0.05)	(0.02)	(0.03)	(0.05)	(0.05)	(0.06)	(0.08)	(0.20)	(0.05)
1000, 1000	0.05	0.00	0.03	0.05	0.05	0.04	0.07	0.17	0.05
	(0.05)	(0.01)	(0.03)	(0.05)	(0.05)	(0.06)	(0.07)	(0.19)	(0.05)
1250, 1250	0.05	0.01	0.03	0.05	0.05	0.02	0.05	0.11	0.05
	(0.05)	(0.02)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)	(0.13)	(0.05)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.01	0.04	0.08	0.05
	(0.05)	(0.02)	(0.03)	(0.05)	(0.05)	(0.02)	(0.04)	(0.09)	(0.05)

Table C.2: False alarm rates and standard errors in the balanced condition with equal sample sizes in reference and focal group.

Unbalanced DIF: Advantage for the reference group

The false alarm rates and hit rates for the unbalanced DIF condition are depicted in Figure 2 in our main article. The corresponding section provides a detailed interpretation of the results. Here, we give a short summary. The false alarm rates are listed together with the standard errors only for equal sample sizes in Table C.4, the hit rates in Table C.5.

hit rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250, 250	0.75	0.44	0.67	0.74	0.73	0.22	0.64	0.66	0.74
	(0.17)	(0.22)	(0.19)	(0.17)	(0.17)	(0.15)	(0.19)	(0.18)	(0.17)
500, 500	0.95	0.81	0.92	0.94	0.94	0.54	0.87	0.86	0.94
	(0.09)	(0.17)	(0.11)	(0.09)	(0.09)	(0.15)	(0.13)	(0.13)	(0.09)
750, 750	0.99	0.95	0.98	0.99	0.99	0.74	0.96	0.94	0.99
	(0.04)	(0.09)	(0.05)	(0.05)	(0.05)	(0.16)	(0.08)	(0.09)	(0.05)
1000, 1000	1.00	0.99	0.99	1.00	1.00	0.88	0.99	0.98	1.00
	(0.02)	(0.05)	(0.03)	(0.02)	(0.02)	(0.13)	(0.05)	(0.05)	(0.03)
1250, 1250	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00
	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.09)	(0.03)	(0.03)	(0.01)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.06)	(0.02)	(0.02)	(0.01)
0.30									
250, 250	0.76	0.43	0.67	0.75	0.74	0.24	0.63	0.65	0.74
	(0.13)	(0.16)	(0.14)	(0.12)	(0.12)	(0.11)	(0.13)	(0.13)	(0.12)
500, 500	0.96	0.82	0.93	0.95	0.95	0.55	0.86	0.84	0.95
	(0.06)	(0.12)	(0.07)	(0.06)	(0.06)	(0.09)	(0.10)	(0.12)	(0.06)
750, 750	0.99	0.95	0.99	0.99	0.99	0.71	0.95	0.92	0.99
	(0.02)	(0.06)	(0.03)	(0.03)	(0.03)	(0.12)	(0.07)	(0.09)	(0.03)
1000, 1000	1.00	0.99	1.00	1.00	1.00	0.85	0.99	0.98	1.00
	(0.01)	(0.03)	(0.02)	(0.01)	(0.01)	(0.12)	(0.03)	(0.05)	(0.01)
1250, 1250	1.00	1.00	1.00	1.00	1.00	0.94	1.00	1.00	1.00
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.08)	(0.02)	(0.02)	(0.01)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.04)	(0.01)	(0.01)	(0.00)
0.45									
250, 250	0.71	0.42	0.63	0.70	0.68	0.26	0.59	0.61	0.69
	(0.10)	(0.13)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
500, 500	0.92	0.79	0.89	0.91	0.90	0.54	0.82	0.77	0.90
	(0.06)	(0.10)	(0.07)	(0.06)	(0.07)	(0.08)	(0.10)	(0.12)	(0.07)
750, 750	0.98	0.93	0.97	0.97	0.97	0.67	0.90	0.86	0.97
	(0.04)	(0.06)	(0.04)	(0.04)	(0.04)	(0.11)	(0.09)	(0.12)	(0.04)
1000, 1000	0.99	0.97	0.99	0.99	0.99	0.81	0.96	0.94	0.99
	(0.02)	(0.04)	(0.03)	(0.02)	(0.03)	(0.13)	(0.05)	(0.09)	(0.03)
1250, 1250	1.00	0.99	1.00	1.00	1.00	0.91	0.99	0.98	1.00
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.09)	(0.03)	(0.05)	(0.02)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.97	0.99	0.99	1.00
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.05)	(0.02)	(0.03)	(0.01)

Table C.3: Hit rates and standard errors in the balanced condition with equal sample sizes in reference and focal group.

Summary

In the unbalanced condition, the SA-selection strategy was superior to the AO-selection strategy when the sample size and the DIF proportion were high as expected, since it not only allowed a higher hit rate but it also displayed a lower false alarm rate.

false alarm rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250, 250	0.07	0.01	0.04	0.06	0.06	0.01	0.07	0.10	0.06
	(0.04)	(0.02)	(0.03)	(0.04)	(0.04)	(0.01)	(0.05)	(0.08)	(0.04)
500, 500	0.09	0.01	0.05	0.06	0.05	0.02	0.09	0.13	0.05
	(0.04)	(0.02)	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.10)	(0.04)
750, 750	0.11	0.02	0.06	0.05	0.05	0.02	0.06	0.13	0.05
	(0.05)	(0.02)	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.10)	(0.04)
1000, 1000	0.13	0.02	0.08	0.05	0.05	0.01	0.04	0.10	0.05
	(0.05)	(0.03)	(0.04)	(0.04)	(0.04)	(0.02)	(0.04)	(0.09)	(0.04)
1250, 1250	0.15	0.03	0.09	0.05	0.05	0.00	0.03	0.07	0.05
	(0.05)	(0.03)	(0.05)	(0.04)	(0.04)	(0.01)	(0.03)	(0.07)	(0.04)
1500, 1500	0.17	0.03	0.10	0.05	0.06	0.00	0.03	0.05	0.05
	(0.05)	(0.03)	(0.05)	(0.04)	(0.04)	(0.01)	(0.03)	(0.05)	(0.04)
0.30									
250, 250	0.13	0.02	0.08	0.09	0.08	0.01	0.10	0.14	0.09
	(0.06)	(0.03)	(0.05)	(0.06)	(0.06)	(0.03)	(0.06)	(0.11)	(0.06)
500, 500	0.21	0.04	0.13	0.09	0.08	0.04	0.15	0.21	0.07
	(0.06)	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)	(0.08)	(0.16)	(0.06)
750, 750	0.29	0.07	0.17	0.09	0.08	0.05	0.13	0.21	0.05
	(0.07)	(0.05)	(0.07)	(0.07)	(0.06)	(0.06)	(0.08)	(0.18)	(0.05)
1000, 1000	0.36	0.09	0.20	0.08	0.09	0.04	0.08	0.17	0.05
	(0.07)	(0.07)	(0.09)	(0.06)	(0.06)	(0.07)	(0.07)	(0.15)	(0.04)
1250, 1250	0.43	0.12	0.23	0.08	0.10	0.02	0.04	0.11	0.05
	(0.07)	(0.08)	(0.10)	(0.06)	(0.06)	(0.05)	(0.05)	(0.12)	(0.04)
1500, 1500	0.50	0.14	0.24	0.07	0.11	0.00	0.03	0.07	0.05
	(0.07)	(0.09)	(0.11)	(0.06)	(0.07)	(0.02)	(0.04)	(0.08)	(0.04)
0.45									
250, 250	0.24	0.05	0.16	0.24	0.23	0.02	0.17	0.22	0.21
	(0.08)	(0.05)	(0.08)	(0.11)	(0.13)	(0.04)	(0.09)	(0.15)	(0.12)
500, 500	0.41	0.12	0.29	0.37	0.34	0.09	0.30	0.38	0.26
	(0.09)	(0.09)	(0.11)	(0.14)	(0.18)	(0.09)	(0.13)	(0.24)	(0.15)
750, 750	0.56	0.19	0.39	0.48	0.40	0.16	0.34	0.46	0.20
	(0.09)	(0.13)	(0.15)	(0.15)	(0.21)	(0.15)	(0.16)	(0.31)	(0.16)
1000, 1000	0.68	0.26	0.47	0.56	0.45	0.19	0.25	0.44	0.11
	(0.08)	(0.19)	(0.18)	(0.16)	(0.23)	(0.20)	(0.18)	(0.36)	(0.12)
1250, 1250	0.77	0.32	0.54	0.66	0.52	0.15	0.13	0.31	0.07
	(0.07)	(0.25)	(0.21)	(0.16)	(0.24)	(0.22)	(0.15)	(0.36)	(0.08)
1500, 1500	0.83	0.38	0.59	0.73	0.58	0.07	0.06	0.16	0.06
	(0.07)	(0.30)	(0.23)	(0.16)	(0.25)	(0.17)	(0.09)	(0.26)	(0.07)

Table C.4: False alarm rates and standard errors in the unbalanced condition with equal sample sizes in reference and focal group.

In the condition of unbalanced DIF, the false alarm rates were no longer well-controlled. When the DIF proportion was high, only the single-anchor-SA, the four-anchor-SA and the iterative-forward-SA method had low false alarm rates in regions of large sample sizes. Both constant single-anchor methods yielded low false alarm rates – but also low hit rates – when the sample size was small.

hit rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250, 250	0.64	0.29	0.53	0.70	0.67	0.07	0.40	0.40	0.71
	(0.19)	(0.19)	(0.20)	(0.20)	(0.19)	(0.11)	(0.20)	(0.23)	(0.20)
500, 500	0.91	0.65	0.85	0.95	0.91	0.28	0.74	0.71	0.95
	(0.11)	(0.21)	(0.14)	(0.09)	(0.10)	(0.23)	(0.19)	(0.22)	(0.09)
750, 750	0.98	0.87	0.96	0.99	0.98	0.59	0.94	0.91	0.99
	(0.06)	(0.14)	(0.08)	(0.04)	(0.06)	(0.28)	(0.10)	(0.14)	(0.03)
1000, 1000	1.00	0.95	0.99	1.00	1.00	0.82	0.99	0.98	1.00
	(0.03)	(0.09)	(0.04)	(0.01)	(0.03)	(0.23)	(0.04)	(0.07)	(0.01)
1250, 1250	1.00	0.99	1.00	1.00	1.00	0.95	1.00	1.00	1.00
	(0.01)	(0.05)	(0.02)	(0.01)	(0.01)	(0.12)	(0.02)	(0.03)	(0.01)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	(0.01)	(0.02)	(0.01)	(0.00)	(0.01)	(0.04)	(0.01)	(0.01)	(0.00)
0.30									
250, 250	0.47	0.16	0.36	0.56	0.55	0.05	0.29	0.30	0.58
	(0.13)	(0.11)	(0.14)	(0.18)	(0.18)	(0.07)	(0.14)	(0.18)	(0.18)
500, 500	0.76	0.40	0.68	0.89	0.85	0.21	0.61	0.58	0.92
	(0.11)	(0.16)	(0.13)	(0.10)	(0.09)	(0.17)	(0.17)	(0.22)	(0.09)
750, 750	0.90	0.63	0.86	0.98	0.93	0.45	0.87	0.82	0.99
	(0.08)	(0.17)	(0.10)	(0.04)	(0.05)	(0.24)	(0.13)	(0.20)	(0.03)
1000, 1000	0.96	0.78	0.95	1.00	0.97	0.71	0.97	0.95	1.00
	(0.05)	(0.15)	(0.06)	(0.02)	(0.04)	(0.24)	(0.06)	(0.12)	(0.01)
1250, 1250	0.99	0.90	0.98	1.00	0.99	0.90	1.00	0.99	1.00
	(0.03)	(0.10)	(0.04)	(0.01)	(0.03)	(0.17)	(0.02)	(0.06)	(0.00)
1500, 1500	1.00	0.95	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	(0.02)	(0.08)	(0.02)	(0.00)	(0.02)	(0.07)	(0.01)	(0.01)	(0.00)
0.45									
250, 250	0.28	0.07	0.19	0.28	0.27	0.03	0.17	0.19	0.32
	(0.09)	(0.06)	(0.09)	(0.13)	(0.14)	(0.04)	(0.10)	(0.14)	(0.15)
500, 500	0.50	0.19	0.41	0.53	0.51	0.11	0.37	0.37	0.63
	(0.10)	(0.12)	(0.13)	(0.15)	(0.19)	(0.12)	(0.16)	(0.24)	(0.18)
750, 750	0.67	0.35	0.61	0.72	0.70	0.25	0.61	0.53	0.87
	(0.09)	(0.18)	(0.16)	(0.14)	(0.19)	(0.23)	(0.21)	(0.32)	(0.13)
1000, 1000	0.78	0.49	0.75	0.83	0.79	0.45	0.85	0.68	0.96
	(0.08)	(0.25)	(0.16)	(0.12)	(0.18)	(0.33)	(0.16)	(0.35)	(0.05)
1250, 1250	0.85	0.61	0.84	0.88	0.81	0.68	0.96	0.82	0.98
	(0.07)	(0.29)	(0.15)	(0.10)	(0.22)	(0.36)	(0.08)	(0.32)	(0.03)
1500, 1500	0.90	0.68	0.90	0.92	0.82	0.87	0.99	0.93	0.99
	(0.06)	(0.32)	(0.12)	(0.09)	(0.25)	(0.26)	(0.03)	(0.22)	(0.02)

Table C.5: Hit rates and standard errors in the unbalanced condition with equal sample sizes in reference and focal group.

All methods from the constant anchor class, especially in regions of small sample sizes, showed poor hit rates. The highest hit rate – in all settings from the unbalanced condition – occurred for the newly proposed iterative-forward-SA method.

D. The impact of anchor contamination

Section 6 in the main article already provided a brief focus on the aspect of anchor contamination. Here, we want to provide a more detailed discussion. As already stated, Figure 3 (top row) in our main article depicts the proportion of replications where at least one item of the anchor was a simulated DIF item (top-left) – this is referred to as *risk of contamination* – and the proportion of simulated DIF items in the anchor when the anchor was contaminated (top-right) – this is referred to as *degree of contamination* together with the false alarm rates (bottom row) including only the replications that resulted in a contaminated anchor (bottom-left) next to those including only the replications that resulted in a pure i.e. DIF-free anchor (bottom-right). If none of these pure replications resulted, the respective false alarm rate is omitted.

The results showed the following: For the all-other method all items functioned as anchor items. Correspondingly, the risk of contamination was 100% and the degree was 45% as simulated. With increasing sample size, the power of detecting artificial DIF (DIF-free items that displayed DIF due to the employed anchor method) increased and, thus, the false alarm rate rose.

Regarding methods from the constant anchor class, the risk of contaminated anchors decreased when the sample size increased for the SA- or the NC-selection strategy, while the AO-selection strategy showed a relatively constant risk of contaminated anchors (observed maximum: four-anchor-AO: 91%, single-anchor-AO method: 40%). If the constant single-anchor items were contaminated, inevitably, the false alarm rates exploded when the sample size was large enough to detect significant artificial DIF (observed maximum: single-anchor-AO: 0.72, single-anchor-SA: 0.52).

Surprisingly, there was a large gap between the degree of contamination for the constant four-anchor methods: When the AO-strategy or the SA-strategy were chosen and the sample size was large, on average about one to one and a half out of four anchor items had DIF. In contrast to this, about three out of four anchor items had DIF for the four-anchor-NC method. In contaminated situations, consequently, the four-anchor-NC method displayed a larger false alarm rate (observed maximum: 0.83) than the four-anchor-AO (observed maximum: 0.65) or the four-anchor-SA method (observed maximum: 0.37). Therefore, the four-anchor-NC method displayed larger false alarm rates compared to the four-anchor-SA method over all unbalanced conditions with 45% DIF items (see again Figure 2 in our main article, top row), even though it had a lower risk of contamination (see again Figure 3 in our main article, top-left). Hence, the degree of contamination was important for the results of the DIF assessment. Note, however, that even if the anchor was pure, the false alarm rates of the constant anchor methods exceeded the significance level. To clarify this fact, we will present an additional simulation study in the next section.

The longer iterative anchors were more often contaminated, as expected (see Figure 3 in our main article). For the iterative-forward-AO method even all replications were contaminated. The iterative-forward-SA and the iterative-backward-AO method yielded a risk of contamination that decreased with the sample size (observed minimum: 0.42 and 0.89). In case of contaminated anchors, the methods from the iterative forward and backward class also produced inflated false alarm rates. When the sample size in each group exceeded 750, the iterative-forward-SA method definitely had the lowest false alarm rate.

Summary

Our findings clarify that it is not the risk of contamination alone that explains the inflated false alarm rates. The best method – in terms of a low false alarm rate together with a high hit rate – in the unbalanced condition when the sample size was large was the iterative-forward-SA method even if it had a high risk of contamination. Therefore, the consequences of contamination depended on the degree of contamination which was low for this method due to the suitable SA-selection strategy. Research on anchor methods should, thus, not only concentrate on the risk of contamination, but also focus on the consequences, which strongly depend on the proportion of contaminated items in the anchor.

E. Characteristics of the anchor items inducing artificial DIF

In this supplement, we provide a more detailed description of the finding from our simulation study that several anchor methods – especially the four-anchor-SA and the four-anchor-NC method – displayed inversely u-shaped false alarm rates.

Our explanation – that the inversely u-shaped pattern results from an interaction between the decreasing extent of artificial DIF induced by anchor contamination and the increasing power of detecting statistically significant artificial DIF – is consistent with the findings from the previous section and with Section 6 of our main article, where the anchor was contaminated: The four-anchor-SA method, for example, displayed a degree of contamination that decreased with sample size (see again Figure 3 in the main article, top-right) and an inversely u-shaped false alarm rate when the anchor was contaminated (see again Figure 3 in the main article, bottom-left).

This situation of contaminated anchors is here addressed in more detail for the constant four-anchor methods (the four-anchor-SA and the four-anchor-NC method that displayed inversely u-shaped patterns as well as the four-anchor-AO method that displayed an increasing false alarm rate, see again Figure 2 in the main article, top-right). In case of contaminated anchors (Figure 4 in the main article, left), the four-anchor methods displayed negative scale shifts. Even though the absolute scale shifts were almost constant over the sample size in regions of small to medium sample sizes for the four-anchor-AO and

four-anchor-NC or even slightly decreasing for the four-anchor-SA method, the false alarm rates rose with growing sample size in the respective range of the sample sizes (Figure 3 in the main article, bottom-left). We attribute this fact to the increasing power of detecting artificial DIF. This also explains the increasing false alarm rates of the four-anchor-AO and the four-anchor-NC methods: The absolute scale shifts were almost constant over the simulated range of the sample size but the false alarm rates increased (Figure 3 in the main article, bottom-left).

For the four-anchor-SA method the absolute scale shift also decreased with increasing sample size in regions of medium or large sample sizes (Figure 4 in our main article, left) and so did the false alarm rate in the respective range of the sample sizes (Figure 3 in our main article, bottom-left).

However, in case of pure replications, the scale shift of the benchmark method was fluctuating around zero, whereas the scale shift of all remaining constant four-anchor methods was negative (Figure 4 in our main article, right) and decreasing with the sample size.

These findings explain why the u-shaped patterns occurred for the four-anchor-SA and the four-anchor-NC method: These methods were able to reduce the absolute scale shift with increasing sample size because the scale shift in pure replications reduced and the risk of contamination reduced as well (i.e. the number of pure replications increased). Taking the increasing power of detecting artificial DIF with growing sample size into account, an inversely u-shaped pattern resulted for the false alarm rates. In contrast to this, the four-anchor-AO method always displayed a relatively high scale shift (that only reduced slightly when the anchor was pure). The power of detecting artificial DIF increased with growing sample size and, therefore, the false alarm rate showed an increase and no considerable decrease.

Summary

In summary, the interaction between a decreasing extent of artificial DIF and an increasing statistical power to detect artificial DIF with growing sample size resulted in an inversely u-shaped false alarm rate. The risk and degree of contamination alone cannot explain the presence of artificial DIF. The anchor items selected by certain anchor selection strategies differed systematically from randomly chosen pure anchor items even if the located anchor items were by definition DIF-free. Counterintuitively instead of items with small differences, these methods tended to select exactly those items with large differences. Therefore, the anchor items found by the SA-, the NC- or the AO-selection strategy displayed a negative scale shift in the additional simulation study and, thus, shifted the scales apart and induced artificial DIF.

This implies that not only the risk and the degree of contamination but also the scale shift in by definition pure replications should be regarded when anchor methods are developed

and investigated in simulation studies. Otherwise, inflated false alarm rates might occur even if the anchor is pure.

F. Summary and discussion

Practical recommendations

Our simulation study highlights the importance of the anchor selection for the correct classification of DIF and DIF-free items and the necessity of a careful consideration of the anchor method to avoid high misclassification rates and doubtful test results.

In case of balanced DIF, the all-other method was slightly better than the iterative-forward-SA strategy. However, due to the fact that the all-other method resulted in seriously inflated false alarm rates when the situation was unbalanced – and that it is doubtful whether the situation of balanced DIF is ever met in practice (Wang and Yeh 2003; Wang *et al.* 2012) – the usage of this anchor method is inadvisable.

Thus, the newly suggested iterative-forward-SA strategy is recommended. When the sample size was large enough, the false alarm rates were low in any condition even if the anchor was contaminated and the hit rates grew rapidly. The adequacy of the selection strategies – by single-anchor (SA) or by all-other (AO) – depended on the DIF situation. In the balanced condition, the AO-selection strategy performed suitable, whereas in the unbalanced condition the SA-selection strategy was more appropriate. But when the iterative-forward class was used, the advance of the AO-selection strategy was marginal. Therefore, we recommend the newly suggested iterative-forward-SA method over the iterative-forward-AO method.

Future research

While our research was limited to DIF detection in the Rasch model using the Wald test, future research may investigate the usefulness of the iterative-forward-SA method for other IRT models and combine it with other DIF detection methods.

In particular, it would be interesting to explore the possibility of employing the iterative-forward strategy together with other IRT-based tests, such as the widely used (see, e.g., Woods 2009; González-Betanzos and Abad 2012) likelihood ratio test, and investigate its compatibility with non-IRT based methods. Future research may e.g. investigate whether those items selected as anchor items by the newly suggested iterative-forward-SA method with the Wald test (or an alternative DIF test) also provide a useful matching criterion for non-IRT based tests. The test results could then be compared with those of classical purification methods that were previously found to improve the final test results (see Miller and Oshima 1992; Clauser, Mazor, and Hambleton 1993; Navas-Ara and Gòmez-Benito 2002, and the references therein).

When other IRT models were the underlying data generating process, previous research found highly discriminating items to be better suited as anchor items (Lopez Rivas, Stark, and Chernyshenko 2009; González-Betanzos and Abad 2012). Thus, the iterative forward procedure might also be combined with a minimum discrimination requirement for the anchor candidates.

Furthermore, the iterative forward anchor class with the SA-selection may be compared with modifications of the anchor selection strategy. For example, Shih and Wang (2009) suggest to use the items corresponding to the lowest rank of the mean absolute DIF statistics similar to the rank-based strategy of Woods (2009). Then items are anchor candidates if they display the lowest mean DIF test statistic when every item is tested for DIF using every other item as constant single-anchor. This modification may be less affected by sample size. Wang *et al.* (2012) established an improvement of the AO-selection strategy by incorporating additional iterations. Firstly, every item is tested for DIF using the all-other method. Then, iteratively, DIF items are excluded from the anchor candidates and a new DIF analysis using the current anchor is conducted until two steps reach the same results. Finally, the anchor items are selected from the remaining candidates using the rank-based strategy. Future research could compare the improved AO-selection to the SA-selection strategy.

Moreover, the DIF test results may also be improved by the construction of new anchor selection strategies. Ideally, the anchor items are DIF-free and induce no artificial scale shift. Furthermore, the impact of the degree of contamination is important for the appropriateness of the results in DIF detection. Therefore, improving the anchor selection strategies with the aim to locate anchors with a small degree of contamination remains an important task.

References

Andrich D, Hagquist C (2012). "Real and Artificial Differential Item Functioning." *Journal of Educational and Behavioral Statistics*, **37**(3), 387–416.

Clauser B, Mazor K, Hambleton RK (1993). "The Effects of Purification of Matching Criterion on the Identification of DIF Using the Mantel-Haenszel Procedure." *Applied Measurement in Education*, **6**(4), 269–279.

González-Betanzos F, Abad FJ (2012). "The Effects of Purification and the Evaluation of Differential Item Functioning with the Likelihood Ratio Test." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **8**(4), 134–145.

Lopez Rivas GE, Stark S, Chernyshenko OS (2009). "The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test." *Applied Psychological Measurement*, **33**(4), 251–265.

Miller MD, Oshima T (1992). "Effect of Sample Size, Number of Biased Items, and Magnitude of Bias on a Two-Stage Item Bias Estimation Method." *Applied Psychological Measurement*, **16**(4), 381–388.

Navas-Ara MJ, Gòmez-Benito J (2002). "Effects of Ability Scale Purification on the Identification of DIF." *European Journal of Psychological Assessment*, **18**(1), 9–15.

Shih CL, Wang WC (2009). "Differential Item Functioning Detection Using the Multiple Indicators, Multiple Causes Method with a Pure Short Anchor." *Applied Psychological Measurement*, **33**(3), 184–199.

Stark S, Chernyshenko OS, Drasgow F (2006). "Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy." *Journal of Applied Psychology*, **91**(6), 1292–1306.

Strobl C, Kopf J, Zeileis A (2010). "Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben." In S Trepte, M Verbeet (eds.), *Wissenswelten des 21. Jahrhunderts – Erkenntnisse aus dem Studentenpisa-Test des SPIEGEL*. VS Verlag, Wiesbaden.

Thissen D (2001). *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. Unpublished manuscript, University of North Carolina, Chapel Hill.

Thissen D, Steinberg L, Wainer H (1988). "Use of Item Response Theory in the Study of Group Differences in Trace Lines." In H Wainer, HI Braun (eds.), *Test Validity*, chapter 10. Lawrence Erlbaum, Hillsdale, New Jersey.

Trepte S, Verbeet M (eds.) (2010). *Allgemeinbildung in Deutschland – Erkenntnisse aus dem SPIEGEL Studentenpisa-Test*. VS Verlag, Wiesbaden.

Wang WC (2004). "Effects of Anchor Item Methods on the Detection of Differential Item Functioning within the Family of Rasch Models." *Journal of Experimental Education*, **72**(3), 221–261.

Wang WC, Shih CL, Sun GW (2012). "The DIF-Free-Then-DIF Strategy for the Assessment of Differential Item Functioning." *Educational and Psychological Measurement*, **72**(4), 687–708.

Wang WC, Yeh YL (2003). "Effects of Anchor Item Methods on Differential Item Functioning Detection with the Likelihood Ratio Test." *Applied Psychological Measurement*, **27**(6), 479–498.

Woods CM (2009). "Empirical Selection of Anchors for Tests of Differential Item Functioning." *Applied Psychological Measurement*, **33**(1), 42–57.

Affiliation

Dr. Julia Kopf
Department of Statistics
Ludwig-Maximilians-Universität München
Ludwigstraße 33
DE-80539 München, Germany
Telephone: +49 89 2180 17146
E-mail: Julia.Kopf@stat.uni-muenchen.de

Prof. Dr. Achim Zeileis
Department of Statistics
Universität Innsbruck
Universitätsstraße 15
AT-6020 Innsbruck, Austria
Telephone: +43 512 507 7103
E-mail: Achim.Zeileis@R-project.org

Prof. Dr. Carolin Strobl
Department of Psychology
Universität Zürich
Binzmühlestrasse 14
CH-8050 Zürich, Switzerland
Telephone: +41 44 63 57370
E-mail: Carolin.Strobl@psychologie.uzh.ch