A note on equating test scores with covariates

Marie Wiberg

Key words: background information, covariates, future challenges

1 Equating test scores with covariates

Test score equating is the statistical process that is used to ensure that test scores from different versions of a test, for example achievement tests are comparable. There exist a number of data collection designs and test equating methods depending on if we have access to common items on the test versions or common test takers. One common data collection design is the equivalent group (EG) design which require the test taker group to be similar even if the test versions are administered at different places or at different time points. Another common data collection design is the non-equivalent groups with anchor test (NEAT) design which require the access to a number of common items (i.e. an anchor test) that has been given to a large number of test takers. The NEAT design is preferable in many test situations. A problem with the NEAT design is that although two groups might be nonequivalent we may not always have access to an anchor test. Recently, a number of ways to circumvent this has been proposed which not only uses test scores but also covariates. The aim of this note is to highlight some future challenges when equating test scores when we have nonequivalent groups and no anchor test.

During 2015 three independent studies were published which all used covariates in test score equating in different ways. In the first study, by Wiberg and Brännberg (2015), covariates were actively used to improve the test equating through a new design called the non-equivalent groups with covariates (NEC) design. In the NEC design, the test takers are categorized through a

1Professor, Department of Statistics, Umeå School of Business and Economics, Umeå University, Sweden
number of covariates that correlate highly with the test scores and these categories are used in a similar way as anchor test scores are used in the NEAT design.

In the second study, by Haberman (2015), pseudo-equivalent groups were constructed by using background information of the test takers. From this information Haberman created weighted samples of test takers which resembled samples from equivalent groups. Linking, which is similar to equating but less restricted, was performed on the weighted samples. In the third study, by Longford (2015), a test score equating method was proposed which was built on causal inference. Especially the ideas of matching with inverse proportional weighting and matched-pairs based on coarsened propensity scored derived from covariates. In all these methods the key ingredient is the use of covariates. It is interesting that all these three studies were published within a few months in the beginning of 2015, without any reference to each other. As equating test scores is an important part of ensuring fairness in standardized achievement tests I expect more research in test score equating with covariates the next following years as there are many equating methods where one could incorporate covariates, including for example item response theory (IRT) equating.

In Sweden, test score equating are used in the Swedish Scholastic Assessment Test (SweSAT) college admissions test. The SweSAT is a paper and pencil test with 160 multiple-choice binary-scored items which consists a quantitative and a verbal section that are equated separately. The test is given twice a year and since 2011 an anchor test is given to a smaller number of test takers. Before the anchor test was introduced, different test taker groups were used with specific values on their covariates to ensure equivalence between test groups when performing equating with an EG design. Some covariates are regularly collected at each test occasion, and it has been shown that education level and age correlate with the test scores and that gender influence the test scores (Bränberg et al., 1990). Although there is now an anchor test for the SweSAT, covariates are still important to facilitate backward comparison of the test results over the years.

2 Future challenges

There are many future challenges which are connected to test equating and the use of covariates. It must however be emphasized that without good covariates we cannot perform a well working equating with covariates. A huge challenge is thus to examine which covariates could be used and how to regularly collect
them. Possible covariates which one could include are for example the test takers response times to the items or the results from other tests. A guideline when choosing covariates could be to examine the correlation between the covariates and the test scores, as covariates which correlate highly with the test scores tend to work well in the equating situation.

There are many equating methods around and some people may argue that we do not need more equating methods. However there might be situations when one need a new equating method. One such possibility could be to examine the possibility to equate a multidimensional test multidimensionally. It is however unclear if we would gain anything compared with equating the different dimensions separately, as a study comparing the use of multidimensional IRT and unidimensional IRT showed only small differences (Wiberg, 2012). It is also possible that for computerized multistage tests one might need a special equating method in the future.

Another huge challenge when equating large scale assessments, is if one actively use covariates as in the NEC design, is how to reach out to the public and explain the use of covariates in the equating of test scores. It is important that a large scale assessment is perceived as fair for all test takers. In Sweden we have a tradition of explaining how everything works, including the equating method of a standardized test, to the public. Thus, in order to use new equating methods in Sweden one has to be prepared to explain its advantages over traditional equating methods. This is interestingly of less importance in for example the USA, where the equating process is typically secret in many large scale assessments as the tests are developed and administered by private companies and not from a governmental institution as in Sweden.

Finally, during the past few years equating packages has been developed for the freeware R (R Development Core Team, 2015). Equating specific packages include for example equate (Albano, 2014), kequate (Andersson et al., 2013) and SNSequate (González, 2014). Although many equating methods are incorporated in these packages there are still many equating methods which one could incorporate into R packages in order to make them more accessible for the users. This is something I hope will happen within a near future. A final reflection is that there are excellent theoretical test equating books around, and thus the focus should be on implementing recent well working equating methods. Initiatives as new equating R packages and the new applied test equating book by González and Wiberg (in preparation) is thus of great importance.
References


