Listening. Learning. Leading.®

Research Report
ETS RR–12-14

# A Note on the Choice of an Anchor Test in Equating

Sandip Sinharay

Shelby Haberman

Paul Holland

Charles Lewis

September 2012

# A Note on the Choice of an Anchor Test in Equating

Sandip Sinharay,[1] Shelby Haberman, Paul Holland, and Charles Lewis
ETS, Princeton, New Jersey

September 2012

**Associate Editors:** James Carlson and Daniel Eignor

**Reviewers:** Neil Dorans and Anne Fitzpatrick

## Abstract

Anchor tests play a key role in test score equating. We attempt to find, through theoretical derivations, an anchor test with optimal item characteristics. The correlation between the scores on a total test and on an anchor test is maximized with respect to the item parameters for data satisfying several item response theory models. Results suggest that under these models, the *minitest*, the traditionally used anchor test, is not optimal with respect to anchor-test-to-total-test correlation; instead, an anchor test with items of medium difficulty, the *miditest*, seems to be the optimum anchor test. This finding agrees with the empirical findings of Sinharay and colleagues that the miditest mostly has higher anchor-test-to-total-test correlation compared to the minitest and mostly performs as well as the minitest in equating.

Key words: equating, minitest, Rasch model

## Acknowledgments

The authors thank Alina von Davier, Neil Dorans, Dan Eignor, and Anne Fitzpatrick for their invaluable advice and Tim Moses for his help with the data sets. The authors gratefully acknowledge the editorial help of Holly Monteith. Any opinions expressed in this publication are those of the authors and not necessarily of ETS.

# 1 Introduction

It is a widely held belief that an anchor test used in test equating should be a representative or a miniature version (i.e., a *minitest*), with respect to both content and statistical characteristics, of the tests being equated (see, e.g., Kolen & Brennan, 2004, p. 19). To ensure statistical representativeness, the usual practice is to make sure that the mean and spread of the item difficulties of the anchor test are roughly equal to those of the tests being equated (see, e.g., Dorans, Kubiak, & Melican, 1998, p. 5).

The requirement that the anchor test be representative of the total tests (i.e., the tests being equated) with respect to content has been shown to be important by Klein and Jarjoura (1985) and Cook and Petersen (1987). Peterson, Marco, and Stewart (1982) demonstrated the importance of having the mean difficulty of the anchor tests close to that of the total tests. However, the literature does not offer any proof of the superiority of an anchor test for which the spread of the item difficulties is representative of the total tests. Furthermore, a minitest has to include very difficult or very easy items to ensure adequate spread of item difficulties, which can be problematic as such items are usually scarce (one reason being that such items often have poor statistical properties, such as low discrimination, and are thrown out of the item pool). An anchor test that relaxes the requirement on the spread of the item difficulties might be more operationally convenient, especially for testing programs using external anchor tests.

Motivated by the preceding, Sinharay and Holland (2006) focused on anchor tests that (a) are content representative, (b) have the same mean difficulty as the total tests, and (c) have spread of item difficulties less than that of the total tests. They defined a *miditest* as an anchor test with a very small spread of item difficulties and a *semi-miditest* as a test with a spread of item difficulty that lies between those of the miditest and the minitest. These anchor tests, especially the semi-miditest, will often be easier to construct operationally than minitests because there is no need to include very difficult or very easy items in them. Sinharay and Holland (2006) cited several works that suggest that the miditest, which has often been referred to as a test with equivalent items (e.g., Tucker, 1946), will be satisfactory with respect to psychometric properties like reliability and validity. Sinharay and Holland (2006), using a number of simulation studies and a real data example, showed that the miditests and semi-miditests have slightly higher anchor-test-to-total-test correlations than the minitests.

In this report, we attempt to find a theoretical explanation for the preceding result regarding the correlation between a test and an anchor test. For a given total test, we attempt to find an anchor test that has the maximum anchor-test-to-total-test correlation. In section 2, we consider the anchor-test-to-total-test correlation when both tests consist of items that satisfy an item response model. An approximation for the correlation is derived that becomes increasingly accurate as the variance of the examinee proficiency, $\theta$, decreases. For the Rasch model, the approximation suggests that the anchor-test-to-total-test correlation is largest if the item difficulties of the anchor test are all equal to the mean examinee proficiency, which happens for a miditest. The same phenomenon occurs for the two-parameter logistic (2PL) model under further restrictions on the item discrimination parameters. In that situation, then, the conventional minitest is not the optimum anchor test in terms of anchor-test-to-total-test correlation. In section 3, we provide a discussion and possibilities for future work.

## 2 A Theoretical Result

Let $X$ denote the score of an individual on a total test with $m$ items, and let $Y$ denote the score on an external anchor test with $n$ items. Let $X = \sum_{i=1}^{m} U_i$, where the item scores $U_i$, $1 \leq i \leq m$, are 0 or 1 and, as is customary in item response theory (IRT), are conditionally independent given a random ability parameter $\theta$ with mean $\mu$ and variance $\sigma^2$. As in standard IRT, the item characteristic function $P_{Xi}$ for $U_i$, $1 \leq i \leq m$, is defined so that the conditional probability that $U_i = 1$ given $\theta$ is $P_{Xi}(\theta)$, and $P_{Xi}$ is a strictly increasing function that is infinitely differentiable. Let $P'_{Xi}$ and $P''_{Xi}$, respectively, denote the first and second derivatives of $P_{Xi}$. It is assumed that $P'_{Xi}$ and $P''_{Xi}$ are uniformly bounded.

Similarly, $Y = \sum_{i=1}^{n} V_i$, where the item scores $V_i$, $1 \leq i \leq n$, are 0 or 1 and are conditionally independent given $\theta$. The item characteristic function $P_{Yi}$ for $V_i$, $1 \leq i \leq n$, is also strictly increasing and infinitely differentiable, with the first two derivatives $P'_{Yi}$ and

$P''_{Yi}$ uniformly bounded.

As in classical test theory, $X = t_X + e_X$, where $t_X$ is the true score and $e_X$ is the error. The true score $t_X$ is $\tau_X(\theta)$, where the test characteristic curve $\tau_X = \ _{i=1}^{m} P_{Xi}$ of $X$ is infinitely differentiable and strictly increasing. The first and second derivatives of $\tau_X$ are $\tau'_X = \ _{i=1}^{m} P'_{Xi}$ and $\tau''_X = \ _{i=1}^{m} P''_{Xi}$, respectively. The error $e_X$ has conditional expectations and conditional variance

$$V_X(\theta) = \sum_{i=1}^{m} P_{Xi}(\theta)[1 - P_{Xi}(\theta)].$$

The first derivative of $V_X$ is

$$V'_X = \sum_{i=1}^{m} P'_{Xi}(1 - 2P_{Xi}),$$

and the second derivative of $V_X$ is

$$V''_X = \sum_{i=1}^{m} [P''_{Xi}(1 - 2P_{Xi}) - 2(P'_{Xi})^2].$$

Similar results hold for $Y$. Given $\theta$, the errors $e_X$ and $e_Y$ are conditionally independent.

The covariance of $X$ and $Y$ is then

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(t_X, t_Y) \\
&= \mathrm{Cov}\left[\tau_X(\theta) - \tau_X(\mu), \tau_Y(\theta) - \tau_Y(\mu)\right] \\
&= E\left\{[\tau_X(\theta) - \tau_X(\mu)][\tau_Y(\theta) - \tau_Y(\mu)]\right\} \\
&\quad - \left\{E\left[\tau_X(\theta)\right] - \tau_X(\mu)\right\}\left\{E\left[\tau_Y(\theta)\right] - \tau_Y(\mu)\right\}. \qquad (1)
\end{aligned}$$

Because

$$\mathrm{Var}(e_X) = E\left[\mathrm{Var}(e_X|\theta)\right] + \mathrm{Var}\left[E(e_X|\theta)\right] = E\left[V_X(\theta)\right] + \mathrm{Var}(0) = E\left[V_X(\theta)\right],$$

the variance of $X$ is given by

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(t_X) + \mathrm{Var}(e_X) \\
&= \mathrm{Var}\left[\tau_X(\theta) - \tau_X(\mu)\right] + \mathrm{Var}(e_X) \\
&= E[\tau_X(\theta) - \tau_X(\mu)]^2 - \left\{E\left[\tau_X(\theta)\right] - \tau_X(\mu)\right\}^2 + E\left[V_X(\theta)\right]. \qquad (2)
\end{aligned}$$

Let the variance $\sigma^2$ decrease to 0. The use of Taylor's theorem leads to the approximation of $\tau_X(\theta)$ by

$$\tau_X(\mu) + (\theta - \mu)\tau'_X(\mu) \tag{3}$$

and of $V_X(\theta)$ by

$$V_X(\theta) = V_X(\mu) + (\theta - \mu)V'_X(\mu). \tag{4}$$

Results similar to Equations 2, 3, and 4 hold for $\mathrm{Var}(Y)$, $\tau_Y(\theta)$, and $V_Y(\theta)$.

Using the preceding approximations on Equations 1, 2, 3, and 4 and similar expressions for $Y$, we obtain

$$\mathrm{Cov}(X, Y)/\sigma^2 \to \tau'_X(\mu)\tau'_Y(\mu),$$

$$\mathrm{Var}(X) \to V_X(\mu),$$

$$\mathrm{Var}(Y) \to V_X(\mu).$$

Thus the correlation of $X$ and $Y$ satisfies

$$\rho(X, Y)/\sigma^2 \to \frac{\tau'_X(\mu)}{[V_X(\mu)]^{1/2}} \frac{\tau'_Y(\mu)}{[V_Y(\mu)]^{1/2}}. \tag{5}$$

The striking feature here is that the maximization of the limit of $\rho(X, Y)/\sigma^2$ for the anchor test does not depend at all on item characteristics of the total test score $X$.

Consider the 2PL model. Let the item discrimination parameters $a_{Xi}$, $1 \le i \le m$, and $a_{Yi}$, $1 \le i \le n$, be positive real numbers, and let the item difficulties $b_{Xi}$, $1 \le i \le m$, and $b_{Yi}$, $1 \le i \le n$, be real. Denote the logistic distribution function at $x$ real by $L(x) = 1/(1 + e^{-x})$. Then $P_{Xi}(\theta) = L(a_{Xi}(\theta - b_{Xi}))$, $1 \le i \le m$, and $P_{Yi}(\theta) = L(a_{Yi}(\theta - b_{Yi}))$, $1 \le i \le n$, so that $P'_{Xi} = a_{Xi}P_{Xi}(1 - P_{Xi})$, $1 \le i \le m$, and $P'_{Yi} = a_{Yi}P_{Yi}(1 - P_{Yi})$, $1 \le i \le n$.

In the special case of the Rasch model with $P_{Xi}(\theta) = L(a(\theta - b_{Xi}))$,

$$\tau'_X(\mu) = a\sum_{i=1}^{n} P_{Xi}(\mu)[1 - P_{Xi}(\mu)] = aV_X(\mu),$$

and similarly, $\tau'_Y(\mu) = aV_Y(\mu)$. Thus, for the Rasch model, Equation 5 suggests that the limit of $\rho(X, Y)/\sigma^2$ is $a^2[V_X(\mu)V_Y(\mu)]^{1/2}$. The maximum value of $V_Y(\mu)$, and hence the maximum value of $\rho(X, Y)$, is achieved if the item difficulty $b_{Yi}$ is $\mu$ for each anchor item $i$, $1 \leq i \leq n$, which happens only when the anchor test is a miditest. That is because $V_Y(\mu)$ is the sum of the terms $P_{Yi}(\mu)[1 - P_{Yi}(\mu)]$, each of which is maximized when $P_{Yi}(\mu) = 0.5$, which happens only when $b_{Yi} = \mu$. In this case, $a[V_Y(\mu)]^{1/2}$ is $an^{1/2}/2$.

The general case of a 2PL model is a bit more complicated; however, it can be shown, using the property that the first derivative of a function is zero at its maximum, that $\tau'_Y(\mu)/[V_Y(\mu)]^{1/2}$ and hence that $\rho(X, Y)$ is maximized for fixed $a_{Yi}$ if each $b_{Yi}$ is $\mu$, which, again, happens only for a miditest. The ratio $\tau'_Y(\mu)/[V_Y(\mu)]^{1/2}$ is then $(2n^{1/2})^{-1} \quad {}_{i=1}^{n} a_{Yi}$.

The challenge in this analysis is that nothing necessarily follows for the case in which the variance of the examinee proficiency $\theta$ is not small. Nonetheless, the analysis does suffice to indicate that it may not always be desirable to have a minitest as an anchor test.

## 3 Discussion and Future Work

In this report, we demonstrate theoretically that the minitest, the most widely used anchor test, may not be the optimum anchor test with respect to the anchor-test-to-total-test correlation. Instead, the results favor a miditest, an anchor test with all items of medium difficulty. Because medium-difficulty items are more easily available than items with extreme difficulty, this result promises to provide test developers with more flexibility when constructing anchor tests.

The suggestion of a number of experts (e.g., Angoff, 1971; Petersen, Kolen, & Hoover, 1989; von Davier, Holland, & Thayer, 2004) that higher anchor-test-to-total-test correlation leads to better equating then implies that an anchor test with items of medium difficulty may lead to better equating. Sinharay and Holland (2007), using analysis of simulated and real data sets, demonstrated that the miditest indeed leads to better equating compared to the minitest under most practical situations. Similar findings were reported by Liu, Sinharay, Holland, Curley, and Feigenbaum (2011) and Liu, Sinharay, Holland, Feigenbaum, and Curley (2011), who compared the equating performances of miditests and minitests using SAT® data sets.

Our results were derived under the assumption that the variance of the examinee proficiency is small; the proof for the general case (i.e., when the variance is not assumed small) could be a topic for future research. In addition, we assumed that the data follow an IRT model in our derivations. It is possible to extend the result to the case in which one does not make any assumption about the data.

## References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225–244.

Dorans, N. J., Kubiak, A., & Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating.* (ETS Statistical Report No. SR-98-02). Princeton, NJ: ETS.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement, 22,* 197–206.

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

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. *Journal of Educational Measurement, 48*, 361–379.

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). New York, NY: Academic Press.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an* *anchor test* (ETS Research Report No. RR-06-04). Princeton, NJ: ETS.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44,* 249–275.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1–13.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating.* New York, NY: Springer.

## Notes

[1] Dr. Sinharay conducted this study and wrote this report while on staff at ETS. He is currently at CTB/McGraw-Hill.