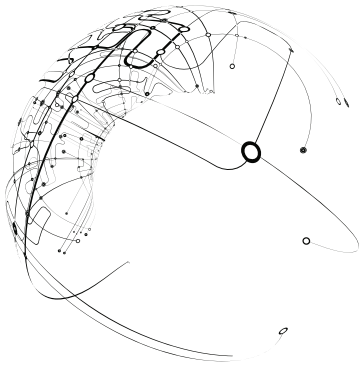

Abstract

Massive open online courses (MOOCs) are playing an increasingly important role in higher education around the world, but despite their popularity, the measurement of student learning in these courses is hampered by cheating and other problems that lead to unfair evaluation of student learning. In this paper, we describe a framework for maintaining test security and preventing one form of cheating in online assessments. We also introduce readers to item response theory, scale linking, and score equating to demonstrate the way these methods can produce fair and equitable test scores. Patrick Meyer is an Assistant Professor in the Curry School of Education at the University of Virginia. He is the inventor of jMetrik, an open source psychometric software program. Shi Zhu is a doctoral student in the Research, Statistics, and Evaluation program in the Curry School of Education. He holds a Ph.D. in History from Nanjing University in China.

AUTHORS

J. Patrick Meyer, Ph.D.
University of Virginia

Shi Zhu, Ph.D.
University of Virginia

CORRESPONDENCE

Email
meyerjp@virginia.edu

Fair and Equitable Measurement of Student Learning in MOOCs: An Introduction to Item Response Theory, Scale Linking, and Score Equating

The last couple of years have witnessed booming development of massive open online courses (MOOCs). These free online courses provide an innovative way of teaching and learning and make higher education accessible to a global audience. Anyone with an internet connection can take courses from top universities in the United States, Canada, Mexico, Europe, Asia, and Australia (Lewin, 2013). MOOCs hold the promise of distributing high quality courses to a global audience and making higher education accessible to people who could not otherwise afford it. Children from working-class families or low-SES backgrounds who could not attend elite universities due to economic reasons are now able to get access to these universities' teaching resources without financial difficulty. Even middle class families can look to MOOCs as a way to offset high tuition rates (Thrift, 2013). Despite the promise of MOOCs, few colleges and universities offer full course credit to students completing a MOOC. Indeed, only five of Coursea's courses are approved for course credit by the American Council on Education (Lederman, 2013), and many professors teaching MOOCs feel that students do not deserve course credit for completing a MOOC (Kolowich, 2013). The concern for course credit not only centers around course quality but also the assessment of student learning.

Online assessments are becoming more important in higher education because students who take online courses do not have many chances to communicate with their instructors and demonstrate mastery of course content in a direct way (Rovai, 2000). One obvious advantage of online assessment over a traditional test is that it can be carried out flexibly in different locations and at different time periods, and can be integrated into the online learning environment (Reeves, 2000). These assessments may simply be online versions of paper-and-pencil tests given in a traditional classroom or they may be innovative

assessments that take full advantage of resources available in computer based testing. For example, personalized e-learning systems based upon item response theory (IRT) can provide adaptive online assessments (Baylari & Montazer, 2009; Chen, Lee, & Chen, 2004) that tailor testing and course content to individual student ability. These adaptive online assessments start with items of moderate difficulty, and then change item difficulty according to a test taker's performance. Given that examinees complete different items and different numbers of items, the final score is not based upon the number of answers he or she got correct but the difficulty and discrimination levels of correctly answered questions (Challis, 2005). Once the score is known, course content can then be tailored to each individual student (see Baylari & Montazer, 2009).

Despite the innovative possibilities with online assessment, there are still some problems that cause concern among educators, policy makers, and test designers such as content disclosure, violations of intellectual property rights, system integrity (Challis, 2005), and identity security (Rovai, 2000). Perhaps the most serious threat to online assessments is cheating, a problem that has long existed in testing.

Despite the innovative possibilities with online assessment, there are still some problems that cause concern among educators, policy makers, and test designers such as content disclosure, violations of intellectual property rights, system integrity, and identity security.

Cizek (1999, 2003) identifies three types of cheating: (a) cheating by giving, taking, or receiving information from others; (b) cheating through use of prohibited materials; and (c) cheating by thwarting the testing process. These types of cheating are observed in traditional paper and pencil testing as well as online testing. Examples of cheating in an online environment include online communication, telecommunication, internet surfing (Rogers, 2006), copying and pasting from online sources (Underwood & Szabo, 2003), obtaining answer keys in an illegitimate way, taking the same assessment several times, and getting unauthorized help during the assessment (Rowe, 2008). Cheating gives dishonest examinees an unfair advantage in the assessment process and it leads assessment professionals to the wrong decision about examinees.

Cohen and Wollack (2006) describe three types of countermeasures that can be used to combat cheating and level the playing field. Human countermeasures require a proctored testing environment and entail any observational methods a test proctor can use to detect cheating. Examples include looking for a student who is conspicuously nervous or who makes frequent trips to the restroom. Electronic countermeasures are similar and may also require a formal testing environment. However, electronic countermeasures make use of technology to prevent and detect cheating. For example, a test environment may use cameras instead of a human proctor to monitor examinees or it may use special equipment to scramble cell phone signals during a test. Electronic countermeasures for an online exam may include installation of security software and IP tracking (Rogers, 2006; Rowe, 2008). Finally, psychometric countermeasures include statistical methods for the prevention and detection of cheating.

Among psychometric counter measures are procedures to limit item exposure (Cohen & Wollack, 2006). If thousands of examinees all complete the same test form, then everyone sees the same items and the risk of an examinee copying and sharing test items with others greatly increases. Prior knowledge of test items will undoubtedly give an advantage to examinees with this information and lead to a breach of standardization procedures and a lack of fairness (Cook & Eignor, 1991). One simple method for reducing item exposure and reducing the impact of cheating is the use of multiple test forms (Cizek, 1999, 2003; Cohen & Wollack, 2006; Cook & Eignor, 1991). This practice reduces exposure and it lessens the possibility that an examinee will cheat because the examinee will not know if the items for which he or she has prior knowledge will actually be on the test he or she is given. Item exposure decreases as the number of test forms increases. In an extreme case, randomly selecting items from a large item pool could result in every examinee completing a unique test form (see Lederman, 2013; Rowe, 2013).

Cook and Eignor (1991) noted that testing must be "fair and equitable" (p. 191). Use of multiple test forms improves fairness by reducing the occurrence of cheating, but it can result in inequities if one test form is easier than another. Students receiving the easier form will perform better and have a greater advantage in seeking MOOC course credit than a student who receives the more difficult form. To achieve Cook and Eignor's fair and equitable criterion, multiple test forms must be placed on a common scale. Scale linking and score

equating results in comparability among scores from different test forms and a situation in which examinees can feel indifferent about the test form they are given. The remainder of this paper discusses the use of item response theory to link and equate multiple test forms. Our discussion focuses on two test forms but it easily extends to any number of test forms or even an entire item bank. As described below, the basic steps in this framework are to: (a) collect data in a way suitable for linking, (b) estimate item and person parameters, (c) link estimates to a common scale, and (d) equate test scores to adjust for test difficulty. The first three steps are required whenever there are multiple test forms. The third step is only needed if the reporting metric is based on the observed score and not the item response theory ability score. Our aim is to introduce readers to this framework. To this end, we have omitted many of the details needed to fully implement this framework.[1]

Item Response Theory

Instructors implicitly rely on classical methods for test scaling and analysis when they create an exam or quiz score by summing the number of items answered correctly by a student. These methods are easy to implement in a classroom setting and provide for well-established methods of analyzing data and evaluating test quality. Tests designed with classical methods give instructors confidence that student scores would not change much if they had given them a different test built to the same content specifications.

Given that examinees complete different items and different numbers of items, the final score is not based upon the number of answers he or she got correct but the difficulty and discrimination levels of correctly answered questions.

Item analysis lies at the heart of evaluating the quality of tests developed through classical methods. Item difficulty and discrimination are two statistics in an item analysis. Item difficulty is the mean item score and item discrimination is the correlation between the item score and test score. These statistics allow instructors to identify problematic items such as those that are too easy or too difficult for students and items that are unrelated to the overall score. Instructors can then improve the measure by revising or eliminating poorly functioning items. An end goal of item analysis is to identify good items and maximize score reliability.

Although classical methods are widely used and easy to implement, they suffer from a number of limitations that are less evident to instructors. One limitation is that classical test theory applies to test scores, not item scores. Item difficulty and discrimination in the classical model are ad hoc statistics that guide test development. They are not parameters in the model. Through rules-of-thumb established through research and practice (see Allen & Yen, 1979), these statistics aid item selection and help optimize reliability. However, they do not quantify the contribution of an individual item to our understanding of the measured trait.

A second limitation to the classical approach is that item statistics and test characteristics are population dependent. Item difficulty will be large (i.e., easier) if a test is given to a group of gifted students, but it will be small (i.e., harder) if the same item is given to a group of academically challenged students. Population effects on item difficulty make it difficult to evaluate item quality because the statistic also reflects examinee quality. Score reliability also depends on the examinee population. It is defined as the ratio of true score variance to observed score variance. As such, scores from a population that is heterogeneous with respect to the measured trait will be more reliable than scores from a population that is homogenous. This result means that an instructor's confidence in the reproducibility of test scores depends on the group of students taking the test (Hambleton & Swaminathan, 1985).

The dependence between item and person characteristics in the classical approach also plays out at the test score level. A test will seem easy if given to a group of gifted students because the average test score will be higher than it is for the general population. Even if multiple test forms are developed to the same specifications and have similar levels of reliability, they will slightly differ in difficulty because of differences in groups taking each form. Equating must be conducted to adjust for these differences and produce comparable scores. Linear and equipercentile equating (see Kolen & Brennan, 2004) are two classical approaches to test equating that use the observed score as the basis of equating.

[1] Readers can find detailed information about test equating in Kolen and Brennan's (2004) *Test equating, scaling and linking: Methods and practices.*

Item response theory (IRT) overcomes these limitations of the classical model. IRT item statistics are estimates of parameters in the model and they can tell us about the contribution of each item to our understanding of the latent trait. Moreover, parameters are invariant to changes in the population, up to a linear transformation (Rupp & Zumbo, 2006). This statement means that if the model fits the data, item parameters will be the same in every population subject to a linear transformation. It also means that person parameters (i.e., the latent trait) will be the same in every group of items that conform to the test specifications (Bond & Fox, 2007). That is, we can obtain the same person ability estimate, within measurement error, from any set of test items. All that we need to do is apply a linear transformation to the parameters from one form to place it on the scale of another. Overcoming the limitations of classical methods does not come without a cost. At a theoretical level, IRT requires more strict assumptions and, at a practical level, it requires more training and specialized software.

Binary Item Response Models

Item response models exist for binary scored (e.g., multiple-choice) and polytomous scored (e.g., constructed response, Likert scales) test questions. For brevity, we will focus on the common unidimensional models for binary items. The most general model is the three parameter logistic (3PL) model. It has one parameter for examinee ability and three parameters for item characteristics. The model is given by

Item difficulty and discrimination in the classical model are ad hoc statistics that guide test development.

$$P(\theta) = c + (1-c)\frac{\exp[a(\theta - b)]}{1+\exp[a(\theta - b)]}.$$

The Greek letter theta, θ, is the examinee ability parameter. It represents a person's latent trait value. The exponential function is indicated by exp in this equation, and the letters *a*, *b,* and *c* represent item parameters.

Item discrimination, the *a* parameter, is the slope of the line tangent to the item characteristic curve (ICC; see Figure 1) at the point of inflection. It reflects the relationship between an item response and the latent trait. It is similar to a factor loading in factor analysis. Item discrimination is always positive. Large item discrimination values will produce an ICC with a steep curve and small values will produce a flat curve. Item difficulty, the *b* parameter, affects the location of the curve. Small difficulty values shift the whole curve to the left and large values shift it to the right. Interpretation of item difficulty in IRT is opposite that for the classical item difficulty statistic, but it is in a more intuitive direction. Small values of item difficulty are easy items, whereas large values are difficult ones. Finally, the guessing parameter, the *c* parameter, indicates the lower asymptote of the ICC. This means that an examinee with an extremely low ability level still has a small chance of answering the item correctly. It is presumed that this small chance is due to guessing on a multiple-choice test.

In the 3PL model, discrimination, difficulty, and guessing can be different for every item. Constraining these parameters leads to different IRT models. The two parameter logistic (2PL) and 1 parameter logistic (1PL) models are special cases of the 3PL. In the 2PL, the guessing parameter is fixed to zero meaning that low ability examinees have a near zero chance of answering the item correctly. The only parameters estimated in the 2PL are item discrimination and difficulty. In the 1PL model, guessing is fixed to zero and discrimination is fixed to be the same for every item but difficulty is freely estimated for every item. That is, discrimination is estimated in the 1PL but a single discrimination value is applied to all items. Item difficulty is also estimated in the 1PL but it is allowed to be different for every item. Finally, the Rasch model is a special version of the 1PL that requires the discrimination parameter to be fixed to a value of one for every item. Only the difficulty parameter is estimated in the Rasch model.

Table 1 lists item parameters for two test forms, Form X and Form Y. However, item parameters are best explained through a graph. An ICC illustrates the probability of a correct answer, $P(\theta)$, for different levels of examinee ability. Figure 1 shows the ICCs for Items 21 and 23 on Form X. As ability increases along the x-axis, the curves increase indicating that the probability of a correct answer increases as the value of the latent trait increases. Item

parameters affect the probability of a correct response and look of the ICC. Item 21 is less discriminating and difficult but involves more guessing than Item 23 (see Table 1). Because of these differences in parameters, the ICC for Item 21 is less steep, shifted to the left, and has a larger lower asymptote than Item 23.

Table 1
Item Parameters for Form X and Form Y before Linking

	Form X Item Parameters			Form Y Item Parameters		
Item	*a*	*b*	*c*	*a*	*b*	*c*
1	**1.17**	**0.56**	**0.11**	**1.31**	**1.09**	**0.10**
2	1.12	-1.39	0.24	1.25	0.35	0.22
3	0.88	-2.40	0.25	1.50	1.48	0.02
4	1.08	-2.87	0.24	1.44	-0.78	0.29
5	**0.95**	**-0.90**	**0.19**	**1.09**	**-0.30**	**0.23**
6	1.01	-0.23	0.19	1.34	-0.30	0.23
7	1.04	1.14	0.02	0.96	0.15	0.25
8	1.15	0.16	0.07	1.22	-0.62	0.22
9	**0.90**	**-0.85**	**0.20**	**1.14**	**-0.01**	**0.26**
10	1.04	0.40	0.23	1.28	-0.44	0.12
11	0.97	-0.24	0.30	1.28	-0.01	0.23
12	1.28	1.16	0.23	1.19	-0.86	0.14
13	**1.07**	**-0.39**	**0.11**	**1.22**	**0.13**	**0.06**
14	1.18	-0.23	0.13	1.26	0.07	0.21
15	0.97	-1.69	0.25	1.39	0.18	0.18
16	1.06	0.94	0.25	1.36	-0.23	0.10
17	**1.27**	**-1.19**	**0.33**	**1.53**	**-0.59**	**0.29**
18	1.29	0.81	0.16	1.06	0.37	0.06
19	0.88	0.90	0.22	0.90	0.46	0.17
20	0.94	-0.33	0.04	1.14	1.46	0.12
21	**0.77**	**-1.26**	**0.19**	**0.95**	**-0.43**	**0.26**
22	0.93	-2.29	0.23	1.08	1.20	0.07
23	1.28	0.25	0.09	1.18	-1.00	0.30
24	1.04	-3.22	0.23	0.98	0.10	0.05
25	**0.96**	**-0.66**	**0.10**	**1.14**	**-0.07**	**0.11**
26	0.93	-1.25	0.23	1.10	0.50	0.10
27	0.98	0.42	0.22	1.04	0.61	0.26
28	0.99	-0.41	0.17	1.10	0.15	0.14
29	**1.14**	**-0.51**	**0.15**	**1.36**	**-0.02**	**0.10**
30	1.00	-0.48	0.26	0.97	-1.07	0.24
Mean	1.04	-0.53	0.19	1.19	0.05	0.17
S.D.	0.13	1.16	0.08	0.17	0.67	0.08

Note: Bold font indicates common items.

Item response theory (IRT) overcomes these limitations of the classical model. IRT item statistics are estimates of parameters in the model and they can tell us about the contribution of each item to our understanding of the latent trait.

Item characteristics in IRT relate directly to test characteristics. A test characteristic curve (TCC) is the sum of all ICCs. It describes the regression of true scores on the latent trait. That is, the x-axis represents person ability, and the y-axis represents true scores. Figure 2 illustrates the TCC for Form X. It looks similar to an ICC but the y-axis is different. The y-axis ranges from the sum of the guessing parameters (5.6 in Figure 2) to the maximum possible sum score (30 in Figure 2). Because of the relationship between a TCC and an ICC, we can select items for a test in a way that achieves a desired TCC.

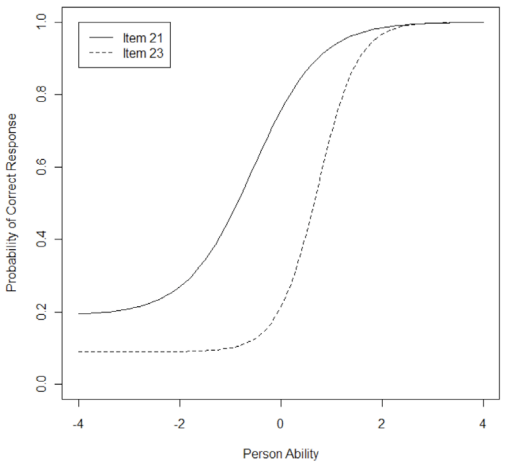

Figure 1. *Item characteristic curves for two items.*

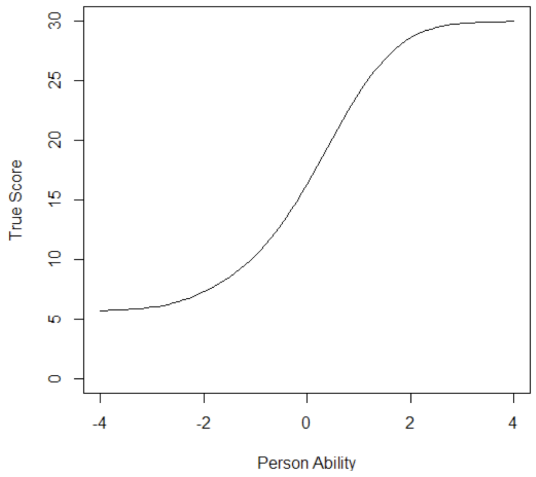

Figure 2. *Test characteristic curve for Form X.*

Another useful function in IRT is the item information function, $I_i(\theta)$. In the 3PL model it is $I_i(\theta) = \left\{a^2[1-P(\theta)]\right\}P^{-1}(\theta)\left\{[P(\theta)-c]^2/[1-c]^2\right\}$. The item information function tells us about the contribution of a single item to our understanding of the latent trait. In the Rasch and 2PL model, information is largest at the place where item difficulty equals examinee ability. Like the ICC, the difficulty parameter affects how far left or right the information curve is shifted and the item discrimination parameter affects how peaked the curve appears. Low difficulty values place information along low levels of the ability scale, whereas larger difficulty values place information at high points of the scale. In a similar vein, large discrimination values concentrate a lot of information over small range of ability levels, but small discrimination values spread a small amount of information over a wide range of the scale. That is, items with large discrimination values tell us more about the latent trait at a particular point than do items with low discrimination. Finally, as the guessing parameter increases, the amount of information decreases.

Figure 3 illustrates the effect of item parameters on the item information function. This figure involves the same two items as Figure 1. The more discriminating item (Item 23) has a more peaked information function than Item 21. It is also shifted to the right because it has a larger difficulty value than Item 21. Notice that these two items tell us very little about examinee ability values less than -2. Most of the information is concentrated between -2.0 and 2.5. To improve information at low ability levels, we should add easier items to the test (e.g., those with a difficulty less than -2.0).

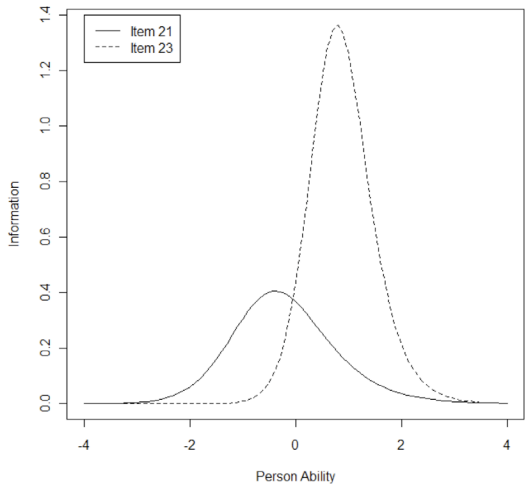

Figure 3. *Item information functions for two items.*

Information also plays a role at the test level. The test information function is the sum of all item information functions. The greater the information, the more we know about the latent trait. Consequently, we can create a test information function that targets specific ability levels, such as the passing score, by selecting items that provide a lot of information at that point. The relationship between item information functions and the test information function make evident the contribution of each item to our understanding of the latent trait. Indeed, information functions are central to many item selection routines in computerized adaptive testing (see Wainer et al., 2000).

Test information is a concept in IRT that replaces the idea of reliability from the classical model in that we aim to maximize information. The reason for maximizing information is because information is inversely related to the standard error of estimating examinee ability, $SE(\theta) = 1/\sqrt{I(\theta)}$. The ability levels with the most information are the ones that have the highest amount of measurement precision.

Parameter Estimation and Software

Marginal maximum likelihood estimation (MMLE) is a method used to obtain parameter estimates in the 2PL and 3PL models. Conditional maximum likelihood (CMLE) and joint maximum likelihood (JMLE) are alternative methods of estimation typically applied to the Rasch family of item response models. For the models discussed in this paper, all of these methods assume we are measuring a single latent trait (unidimensionality) and that items are independent at a given value of the latent trait (conditional independence; see Hambleton & Swaminathan, 1985). We will not discuss the details of these estimation methods, but on a practical level, these methods are synonymous with different types of IRT software. Programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), ICL (Hanson, 2002), and PARSCALE (Muraki & Bock, 1997) offer MMLE for 2PL, 3PL, and polytomous response models. WINSTEPS (Linacre, 2011) and jMetrik (Meyer, 2013) provide JMLE for Rasch family models, and the eRM (Mair & Hatzinger, 2007) package in R provides CML for Rasch family models.

We can create a test information function that targets specific ability levels, such as the passing score, by selecting items that provide a lot of information at that point.

Sample size requirements are another practical consideration for IRT. As a rule of thumb, the more parameters in the model, the larger the sample size that is needed to obtain stable parameter estimates. Rasch models require as little as 100 examinees (Wang & Chen, 2005), but the 3PL model may require at least 1,500 (Mislevy & Stocking, 1989). These sample size requirements are prohibitive for small classrooms and they are one reason why IRT is not used very often in traditional course settings. MOOCs, on the other hand, enroll tens of thousands of students, which is more than enough to obtain accurate estimates with any IRT model. Large class sizes are one reason why IRT and MOOCs are the perfect marriage.

Scale Linking in Item Response Theory

Data must be collected in a particular way in order to implement scale linking. In an equivalent groups design, each test form is given to a random sample of examinees. Items can be completely unique to each test form because the groups are randomly equivalent; test forms are considered to be the only reason for difference in test performance. Consequently, person ability estimates form the basis of scale transformation coefficients that place each form on a common scale.

A popular alternative to the equivalent groups design is the common item nonequivalent groups design (Kolen & Brennan, 1987). In this design, two different groups receive a different test form. For example, one group receives Form X and another group receives Form Y. Each test form includes a set of items unique to the form and a set of items common to both forms. Examinees are considered to be the only reason for differences in performance and parameter estimates for the common items form the basis of scale transformation coefficients. This design is easy to implement in practice but it requires great care in creating the set of common items that are embedded on each test form.

Overall, each form is designed to measure the same content and adhere to the same test specifications. Common items embedded in each form are selected to be a mini or midi version of the complete test and they are placed in about the same position on each form (Kolen & Brennan, 2004; Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011). In a mini version of the test, common items cover the same range of difficulty values as the complete test, and in a midi version, common items cover a narrower range of difficulty. Table 1 demonstrates a common item design with item parameters from two different forms. The items in bold are the items shared by both forms. Once we collect data we can estimate parameters and place both forms on the same scale.

As noted earlier, parameters in an IRT model are invariant up to a linear transformation. If you apply a linear transformation to the person ability parameter and the same transformation to the item parameters, the probability of a correct response remains the same as it was prior to any transformation. This implies that there are no unique parameter values that determine the scale; any linear transformation of the parameters would result in the same probabilities. This problem is referred to as scale indeterminacy and it is resolved in practice by arbitrarily setting the person ability scale to have a mean of zero and a standard deviation of one during

the estimation process. A consequence of resolving scale indeterminacy in this way is that an item that is included on two different test forms will have different parameter estimates. However, we can use the differences in item parameter estimates from both forms to identify the linear transformation that places both forms on the same scale.

Steps for linking test forms to a common scale differ depending on whether estimation is conducted concurrently or separately. In concurrent calibration, data from all test forms are combined into a single data set and the parameters are estimated simultaneously. The overlap in common items will result in estimates that are on a common scale. No further work is needed to place Form X parameters on the scale of Form Y. It is handled automatically during estimation. Fixed common item calibration is a slight variation of this procedure that also places parameters on a common scale during the estimation routine. In this procedure, common item parameters on Form X are fixed to their estimated values on Form Y.

In separate calibration, parameters for each form are estimated separately and an additional step is needed to link estimates to a common scale. A consequence of setting the mean person ability to zero and standard deviation to one during separate estimation of Form X and Form Y parameters is that examinees taking Form X will have the same mean ability level as those taking Form Y even though the two groups may not be equivalent. That is, we end up with within group scales. To adjust the Form X parameters, we use the linear transformation $\theta_{Y*} = A\theta_X + B$ to place a Form X ability, θ_X, on the scale of Form Y. Similar transformations are applied to the item parameters. Discrimination is transformed by $a_{Y*} = a_X / A$ and difficulty is transformed by $b_{Y*} = Ab_X + B$ where the items parameters with an X subscript are parameters that belong to Form X. A and B are transformation coefficients derived from the common item parameters, and there are four popular methods for computing them (Hanson & Béguin, 2002).

MOOCs, on the other hand, enroll tens of thousands of students, which is more than enough to obtain accurate estimates with any IRT model. Large class sizes are one reason why IRT and MOOCs are the perfect marriage.

The mean/sigma (Loyd & Hoover, 1980) and mean/mean (Marco, 1977) methods are referred to as method of moments procedures because they use only item parameter descriptive statistics to compute the transformation coefficients. They are easy to implement and can be computed by hand. For example, mean/sigma transformation coefficients can be computed from the summary statistics in Table 2. The slope coefficient is computed from the common item estimates by dividing the standard deviation of Form Y item difficulty by the standard deviation of Form X item difficulty $A = \sigma(b_Y) / \sigma(b_X)$. The intercept coefficient is the mean item difficulty of Form Y subtracted by the rescaled Form X mean item difficulty, $B = \mu(b_Y) - A\mu(b_X)$. Using Table 2, these coefficients are $A = 0.51 / 0.58 = 0.88$ and $B = -0.02 - 0.88(-0.65) = 0.55$. The slope coefficient differs slightly from the value reported for the mean/sigma method in Table 2 because of rounding. The values in Table 2 are more accurate. The mean/sigma method gets its name because it uses the mean and standard deviation of item difficulty parameters. The mean/mean method, on the other hand, only uses the item discrimination and item difficulty means. It does not involve the computation of standard deviations. Specifically, the slope coefficient for the mean/mean method is $A = \mu(a_X) / \mu(a_Y)$. The intercept is computed in the same way as in the mean/sigma method. Using the values in table 2, the slope is $A = 1.03 / 1.22 = 0.84$ and the intercept is $B = -0.02 - 0.84(-0.65) = 0.53$. These values are slightly different from the tabled values due to rounding.

Table 2
Common Item Descriptive Statistics and Transformation Coefficients

	Form X Item Parameters			Form Y Item Parameters		
Statistic	*a*	*b*	*c*	*a*	*b*	*c*
Mean	1.03	-0.65	0.17	1.22	-0.02	0.17
S.D.	0.16	0.58	0.08	0.18	0.51	0.09

Method	*A*	*B*
Mean/sigma	0.89	0.55
Mean/mean	0.84	0.52
Haebara	0.87	0.53
Stocking-Lord	0.85	0.52

Note: transformation coefficients computed at full precision. Computing coefficients for the moment methods using descriptive statistics in the table above may differ slightly due to rounding.

Method of moments procedures are attractive because of their simplicity, but their main limitations are that they do not use all of the item characteristics and they can be affected by outliers. Alternatively, the Haebara (Haebara, 1980) and Stocking-Lord procedures (Stocking & Lord, 1983) are referred to as characteristic curve methods because they use item and test characteristic curves to obtain the transformation coefficients. Characteristic curve methods are computer intensive and require specialized computer software such as STUIRT (Kim & Kolen, 2004), the plink package in R (Weeks, 2011), and jMetrik (Meyer, 2013). Stocking-Lord and Haebara transformation coefficients are listed in Table 2. We used coefficients from the Stocking-Lord procedure to transform Form X parameters to the scale of Form Y (see Table 3). Parameters estimates in Table 3 are now on a common scale.

Table 3
Item Parameters after Linking with Coefficients from the Stocking-Lord Procedure

	Form X Item Parameters			Form Y Item Parameters		
Item	*a*	*b*	*c*	*a*	*b*	*c*
1	**1.38**	**1.00**	**0.11**	**1.31**	**1.09**	**0.10**
2	1.32	-0.66	0.24	1.25	0.35	0.22
3	1.03	-1.52	0.25	1.50	1.48	0.02
4	1.27	-1.92	0.24	1.44	-0.78	0.29
5	**1.12**	**-0.25**	**0.19**	**1.09**	**-0.30**	**0.23**
6	1.19	0.33	0.19	1.34	-0.30	0.23
7	1.22	1.49	0.02	0.96	0.15	0.25
8	1.36	0.66	0.07	1.22	-0.62	0.22
9	**1.06**	**-0.21**	**0.20**	**1.14**	**-0.01**	**0.26**
10	1.22	0.86	0.23	1.28	-0.44	0.12
11	1.14	0.31	0.30	1.28	-0.01	0.23
12	1.51	1.50	0.23	1.19	-0.86	0.14
13	**1.25**	**0.19**	**0.11**	**1.22**	**0.13**	**0.06**
14	1.38	0.32	0.13	1.26	0.07	0.21
15	1.14	-0.92	0.25	1.39	0.18	0.18
16	1.24	1.32	0.25	1.36	-0.23	0.10
17	**1.50**	**-0.49**	**0.33**	**1.53**	**-0.59**	**0.29**
18	1.52	1.21	0.16	1.06	0.37	0.06
19	1.03	1.29	0.22	0.90	0.46	0.17
20	1.11	0.24	0.04	1.14	1.46	0.12
21	**0.90**	**-0.55**	**0.19**	**0.95**	**-0.43**	**0.26**
22	1.09	-1.42	0.23	1.08	1.20	0.07
23	1.50	0.73	0.09	1.18	-1.00	0.30
24	1.22	-2.21	0.23	0.98	0.10	0.05
25	**1.13**	**-0.04**	**0.10**	**1.14**	**-0.07**	**0.11**
26	1.10	-0.54	0.23	1.10	0.50	0.10
27	1.15	0.88	0.22	1.04	0.61	0.26
28	1.17	0.17	0.17	1.10	0.15	0.14
29	**1.34**	**0.09**	**0.15**	**1.36**	**-0.02**	**0.10**
30	1.17	0.11	0.26	0.97	-1.07	0.24
Mean	1.23	0.07	0.19	1.19	0.05	0.17
S.D.	0.16	0.99	0.08	0.17	0.67	0.08

Note: Bold font indicates common items.

Despite the increasing impact of MOOCs on higher education, cheating poses a threat to online assessments in these courses.

Among the various methods for scale linking, the Stocking-Lord procedure works best when items are all of the same type (Baker & Al-Karni, 1991; Wells, Subkoviak, & Serlin, 2002), and the Haebara method works best in mixed format tests such as those that combine multiple-choice and short answer type items (Kim & Lee, 2006). Concurrent calibration and fixed common item procedures also work very well, particularly compared to the method of moments procedures. However, these two methods make it difficult to detect items that have an undue influence on linking process.

Score Equating with Item Response Theory

Testing programs report scores to examinees in a scaled score metric that is usually limited to positive whole numbers. For example, the GRE Verbal Reasoning scaled score consists of one point increments between 130 and 170. The purpose of scaled scores is to distinguish them from simple sum scores and have a metric that is independent of test forms. They are obtained by either transforming an examinee's IRT ability estimate or an examinee's sum score. In the former case, no further work is needed to produce comparable scaled scores; the linking process has already adjusted for difference among test forms and placed parameters on a common scale. IRT ability parameters are simply transformed to the scaled score and the work is done.

Recall that IRT ability parameters are invariant in different samples of items but observed scores are not. As such, if the scaled score scale is defined as a transformation of the observed score, an additional equating step is needed to adjust test forms for differences in difficulty. True score equating and observed score equating are two options in an IRT framework. True score equating is easier to implement as it involves test characteristic curves from two forms. As illustrated in Figure 4, Form X is easier than Form Y at low levels of ability, but at high levels of ability, the opposite is true. To adjust for these differences in test difficulty, we find the Form Y equivalent of a Form X score. As illustrated by arrows in Figure 4, the steps involve (a) choosing a Form X true score value (21 in Figure 4), (b) finding the Form X ability level that corresponds to that true score (0.61 in Figure 4), (c) computing the Form Y true score at the Form X ability level (22.3 in Figure 4). Thus, a Form X true score of 21 is equivalent to a rounded Form Y true score of 22.

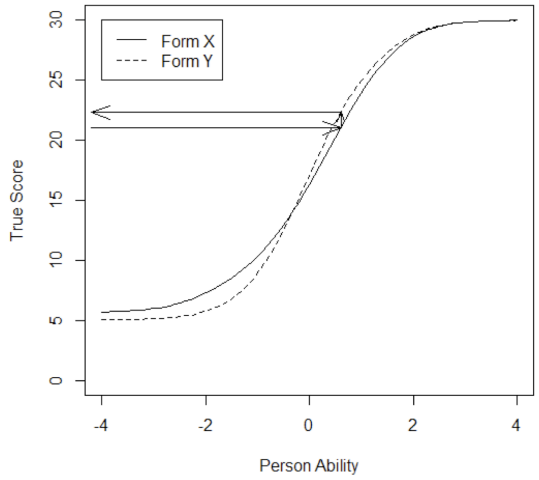

Figure 4. *An illustration of true score equating.*

Although true score equating is easy to illustrate, it actually requires computer intensive methods to implement. POLYEQUATE (Kolen & Cui, 2004), plink (Weeks, 2011), and jMetrik (Meyer, 2013) are three free programs that implement true score equating. Table 4 lists all of the equated true score values for Form X and Form Y. Scores from the two different test forms are now comparable. They have the same meaning and lead to fair and equitable decisions about student performance.

Discussion

Despite the increasing impact of MOOCs on higher education, cheating poses a threat to online assessments in these courses. Students may get illicit help via communication devices or even get access to answers before the assessment. Multiple test forms and extensive item pools can improve test security and increase fairness in online testing, but they leave open the possibility that test forms will differ in difficulty and give an advantage to students completing the easier form. Scale linking and score equating procedures must accompany the use of multiple test forms to ensure comparability among scores. Classical test theory methods commonly used in traditional course assessment can be extended to classical methods of score equating. However, these methods suffer from limitations such as population dependence. Large class sizes that are typical for MOOCs make a wide range of IRT models available for online assessment. IRT based scale linking and score equating overcome many of the problems with classical methods and make scale linking and score equating relatively easy to implement in practice.

Using a single counter–measure to combat cheating is like protecting your home from burglary by locking the doors and leaving the windows open.

Multiple test forms prevent unfair advantages due to prior knowledge of test items and the sharing of answer keys, but they do not prevent all forms of cheating. Indeed, using a single countermeasure to combat cheating is like protecting your home from burglary by locking the doors and leaving the windows open. An effective testing program makes use of multiple countermeasures to address all points of vulnerability. Multiple test forms should be combined

Table 4
True Score Equating Results

Form X True Score	Form X Theta	Form Y True Score Equivalent	Rounded Equivalent
0	-99.0	0.00	0
1	-99.0	0.89	1
2	-99.0	1.78	2
3	-99.0	2.67	3
4	-99.0	3.56	4
5	-99.0	4.45	4
6	-2.98	5.14	5
7	-2.15	5.58	6
8	-1.70	6.28	6
9	-1.36	7.26	7
10	-1.08	8.46	8
11	-0.86	9.79	10
12	-0.66	11.19	11
13	-0.49	12.60	13
14	-0.33	13.99	14
15	-0.18	15.35	15
16	-0.04	16.66	17
17	0.09	17.91	18
18	0.22	19.10	19
19	0.35	20.23	20
20	0.48	21.29	21
21	**0.61**	**22.30**	**22**
22	0.74	23.25	23
23	0.87	24.15	24
24	1.02	25.01	25
25	1.17	25.84	26
26	1.34	26.66	27
27	1.54	27.46	27
28	1.79	28.28	28
29	2.17	29.10	29
30	99.00	30.00	30

Note: Bold font indicates the true score equating relationship illustrated with arrows in Figure 4.

Accomplishing this criterion in practice will drive more institutions to offer course credit for MOOC completion and further expand the influence of these courses on higher education throughout the world.

with other counter measures such as proctored testing to combat cheating in a comprehensive way. Fair and equitable testing is achieved by minimizing all forms of cheating and ensuring the comparability of scores from different test forms. Accomplishing this criterion in practice will drive more institutions to offer course credit for MOOC completion and further expand the influence of these courses on higher education throughout the world.

Limitations

We simulated the data in this paper using a 3PL model. We obtained parameter estimates reported in the tables with ICL (Hanson, 2002) and conducted the linking and equating procedures in jMetrik (Meyer, 2013). We used simulated data to demonstrate IRT, scale linking, and score equating. As such, the data perfectly fit the 3PL model and are void of the usual noise of real test data. Our data also make it appear that equating does not change scores by much. However, this result is not always the case. Scores could change more substantially with real test data and greater difference in test forms. However, the only way to know the extent of the change in scores is to conduct the complete linking and equating process.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147-162.

Baylari, A., & Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications, 36*(4), 8013-8021.

Bond, T.G., & Fox, Ch.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Challis, D. (2005). Committing to quality learning through adaptive online assessment. *Assessment & Evaluation in Higher Education, 30*(5), 519-527.

Chen, C., Lee, H., & Chen, Y. (2005). Personalized e-learning system using item response theory. *Computers & Education, 44*(3), 237-255.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement,* 4th Edition (pp. 355-386). Westport, CT: Praeger.

Cook, L. L., & Eignor, D. R. (1991). An NCME module on IRT Equating methods. *Educational Measurement: Issues and Practice, 10*(3), 191-199.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2003). *Detecting and preventing classroom cheating: Promoting integrity in assessment.* Thousand Oaks, CA: Corwin Press.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hanson, B. A. (2002). *IRT command language* [computer software]. Retrieved from http://www.b-a-h.com/.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* [computer software]. Retrieved from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx

Kim, S., & Lee, W. -C. (2006). An extension of four IRT linking methods for mixed format tests. *Journal of Educational Measurement, 43*, 53-76.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*(3), 263-277.

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

Kolen, M. H., & Cui, Z. (2004). POLYEQUATE [computer software]. Retrieved from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx.

Kolowich, S. (2013, March). The professors who make the MOOCs. *Chronicle of Higher Education.* Retrieved March 20, 2013, from http://chronicle.com/article/The-Professors-Behind-the-MOOC/137905/#id=overview.

Lederman, D. (2013, February). Expanding pathways to MOOC credit. *Inside Higher Education*. Retrieved from http://www.insidehighered.com/news/2013/02/07/ace-deems-5-massive-open-courses-worthy-credit.

Lewin, T. (2013, February 20). Universities abroad join partnerships on the Web. *New York Times.* Retrieved from http://www.nytimes.com/2013/02/21/education/universities-abroad-join-mooc-course-projects.htm

Linacre, J. M. (2011). *Winsteps*® (Version 3.71.0) [computer software]. Beaverton, OR: Winsteps.com.

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement, 48*(4), 361-379.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20. Retrieved from http://www.jstatsoft.org/v20/i09/

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

Meyer, J.P. (2013). *jMetrik version 3* [computer software]. Retrieved from www.ItemAnalysis.com.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Muraki, E., & Bock, R. D. (1997). *PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales* [computer software]. Chicago, IL: Scientific Software International.

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research, 23*(1), 101-111.

Rogers, C. F. (2006). Faculty perceptions about e-cheating during online testing. *Journal of Computing Sciences in Colleges, 22*(2), 206-212.

Rovai, A. P. (2000). Online and traditional assessments: What is the difference? *Internet and Higher Education, 3*(3), 141-151.

Rowe, N. C. (2004). Cheating in online student assessment: Beyond plagiarism. *Online Journal of Distance Learning Administration, 7*(2).

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory* [computer software]. Chicago, IL: Scientific Software International.

Thrift, N. (2013, February 13). To MOOC or not to MOOC. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/blogs/worldwise/to-mooc-or-not-to-mooc/31721

Underwood, J., & Szabo, A. (2003). Academic offences and e-learning: Individual propensities in cheating. *British Journal of Educational Technology, 34*(4), 467-477.

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Weeks, J. P. (2011). *Plink package* [computer software]. Retrieved from http://cran.r-project.org/web/packages/plink/index.html

Wainer, H., Dorans, N. J., Flaughter, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer,* 2nd Edition. Mahwah, NJ: Lawrence Erlbaum.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77-87.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [computer software]. Chicago, IL: Scientific Software International.