applied sciences

MDPI

Article

The Accurate Measurement of Students' Learning in E-Learning Environments

Sunbok Lee [1,*], Youn-Jeng Choi [1] and Hyun-Song Kim [2]

[1] Department of Education, Ewha Womans University, Seoul 03760, Korea; younjengchoi@ewha.ac.kr
[2] Teaching and Learning with Technology, Rutgers University, New Brunswick, NJ 08901, USA; hsk74@docs.rutgers.edu
* Correspondence: sunboklee@ewha.ac.kr; Tel.: +82-2-3277-3338

Abstract: The ultimate goal of E-learning environments is to improve students' learning. To achieve that goal, it is crucial to accurately measure students' learning. In the field of educational measurement, it is well known that the key issue in the measurement of learning is to place test scores on a common metric. Despite the crucial role of a common metric in the measurement of learning, however, less attention has been paid to this important issue in E-learning studies. In this study, we propose to use fixed-parameter calibration (FPC) in an item response theory (IRT) framework to set up a common metric in E-learning environments. To demonstrate FPC, we used the data from the MOOC "Introduction to Psychology as a Science" offered through Coursera collaboratively by Georgia Institute of Technology (GIT) and Carnegie Mellon University (CMU) in 2013. Our analysis showed that the students' learning gains were substantially different with and without FPC.

Keywords: massive open online courses; equating; common metric; fixed-parameter calibration; item response theory

Citation: Lee, S.; Choi, Y.-J.; Kim, H.-S. The Accurate Measurement of Students' Learning in E-Learning Environments. *Appl. Sci.* **2021**, *11*, 9946. https://doi.org/10.3390/app11219946

Academic Editor: Antonio Sarasa Cabezuelo

Received: 1 September 2021
Accepted: 13 October 2021
Published: 25 October 2021

1. Introduction

As an E-learning environment, massive open online courses (MOOCs) are receiving growing attention from both academics and the public for their potential to provide learning opportunities for a large number of students. Such attention has led to greater research interest as well [1,2]. One of the ultimate goals in MOOCs research is to identify factors that can promote students' learning [3]. For examples, previous studies have examined how students' motivation [4], rapid peer feedback [5], and automated personalized feedback [6] influence students' learning in MOOCs. To fairly evaluate the effects of various educational factors on students' learning in MOOCs, it is crucial to precisely measure students' learning [7].

From the measurement point of view, the key issues in the measurement of learning are to use the scores that are independent of measurement instruments and to place the scores on a common metric. These issues are important because any measurement of learning based on scores that are not independent of the measurement instruments and/or are not placed on a common metric will lead to inaccurate estimates of learning and therefore wrong conclusions. To be more concrete, let us consider a typical case in MOOCs where students' learning is operationalized as their gain scores between pre- and post-tests, and let us further assume that the items (or problems) in the post-test are relatively easier than those in the pre-test. In this hypothetical scenario, the fact that a student's proportions of correct answers were 0.6 on the pre-test and 0.8 on the post-test does not necessarily mean that the student shows learning gain because the student's higher proportion of correct answers on the post-test (i.e., 0.8) may be due to the easier items in the post-test, as well as the student's actual learning. This example clearly shows that the proportions of correct answers are not comparable across pre- and post-tests because of the confounding between the students' ability and the items' difficulty.

Another characteristic of MOOCs also requires the establishment of a common metric. Unlike students in the traditional test setting who need solve the same set of items, students in MOOCs usually solve different sets of items because of the self-paced nature of MOOCs. For the same reason as the example of the pre- and post-tests, a common metric needs to be established for any meaningful comparison between students who solved different sets of items in MOOCs. Despite its crucial role, however, less attention has been paid to establishing a common metric in the measurement of learning in MOOCs. Although we illustrated the issue in the specific context of MOOCs, any E-learning environment shares the same problem. In this study, we propose to use the fixed-parameter calibration method (FPC) [8] in item response theory (IRT) to establish a common metric for the accurate measurement of students' learning in E-learning environments.

As indicated by previous examples, the key problem in the measurement of learning in E-learning environments can be formulated as the problem of establishing a common metric across repeated measures over time (e.g., pre- and post-tests) or across different forms of the same test (e.g., different sets of items attempted by different students). The methods for establishing a common metric in such cases have been extensively studied in the field of educational measurement [9]. In the educational measurement literature, a family of procedures for aligning different metrics and making scores comparable is often called linking or equating [10]. Among various proposed procedures, FPC in IRT is expected to establish a common metric in a way that fits well in MOOCs settings.

This article is organized as follows. We first discuss why a common metric is necessary in E-learning environments using the specific example of MOOCs. Then, we present a brief overview of IRT with emphasis on why IRT is a preferable method for the measurement of learning in E-learning environments. Lastly, we present an analysis with real data contrasting students' learning gains based on proportion scores and IRT scores with FPC.

2. Background

2.1. The Needs for a Common Metric in MOOCs

In MOOCs, a common metric is necessary for several reasons. First, a common metric allows us to measure students' learning properly. Student learning is typically operationalized as the gain scores between pre- and post-tests [11]. Researchers may use the same set of items for both the pre- and post-tests. However, in such a case, students are exposed to the same set of items, and therefore, practice effects may be confounded with students' actual learning. When different sets of items are used in the pre- and post-tests, a common metric needs to be established between the pre- and post-test scores. Without a common metric, scores from the pre- and post-tests are not comparable, and therefore, learning gain cannot be properly measured.

Second, a common metric allows us to evaluate students fairly. In MOOCs, students can freely navigate a MOOCs course and can design their own learning process. As a result, students in a MOOC are likely to solve different sets of items across a course because of the self-paced nature of MOOCs [12]. For fair evaluations, therefore, it needs to be considered that students might solve different sets of items or different forms. For example, in many MOOCs, students can obtain certificates if they meet certain criteria on their performance in the course (say a 70% proportion correct on homework, quizzes, and exams). Given the fact that students may solve different sets of items, it might not be fair to use percentage scores for decisions about certificates because percentage scores do not reflect different levels of item difficulties that students encounter on homework, quizzes, and tests.

Third, a common metric allows us to reuse MOOCs items more efficiently. Since MOOCs started, a considerable number of items (or learning resources in general) have been created. Therefore, researchers and practitioners began to consider how to reuse existing learning resources. For example, the edX MOOCs platform provides a custom course on the edX (CCX) with which high school teachers can create their own MOOCs by editing the learning resources in the existing edX MOOCs [13], and the Digital Assets for Reuse in Teaching (DART) project from Harvard university allows any instructor at

Harvard to add edX resources to his/her residential courses in the Canvas Learning Management System (LMS). Item parameters, such as difficulty and discrimination parameters, calibrated on a common metric can maximize the benefit of reusing items by allowing us to customize a set of items that are specifically designed to achieve a certain level of difficulty and discrimination.

Fourth, a common metric allows us to prevent cheating in MOOCs by using multiple test forms. Cheating is a long-lasting problem in testing. In an online environment, cheating occurs in many different ways: online communication [14], copying and pasting from online sources [15], illegitimately obtained answer keys [16], and so on. The use of multiple test forms can reduce cheating and therefore can enhance fairness. A prerequisite for using multiple test forms is to establish a common metric across multiple test forms to make scores from different forms comparable.

2.2. Item Response Theory

IRT is a system of measurement models that has been widely used in educational measurement [17]. In general, measurement models are statistical models that establish the correspondence between latent variables (e.g., students' ability) and their manifestations (e.g., students' responses to items). The ultimate interest in measurement models is to make an inference about unobservable latent variables using observed variables. In essence, IRT is the logistic regression of observed item responses on the person characteristics (i.e., ability) and item characteristics (e.g., difficulty and discrimination). For example, the two-parameter logistic (2PL) model, which is one of the popular IRT models, can be expressed as the following equation [18]:

$$P[Y_i = 1] = \frac{exp[a_i(\theta_j - b_i)]}{1 + exp[a_i(\theta_j - b_i)]},$$

where $P[Y_i = 1]$ represents the probability of getting an item i correct, θ_j represents the ability of a person j, b_i represents the difficulty of an item i, and a_i represents the discrimination of an item i. The equation for the 2PL model can be plotted to describe the relationship between $P[Y_i = 1]$ and ability θ_j, which is often called an item characteristic curve (ICC). Figure 1 presents the ICCs for three different sets of item parameters. From the comparison between Line 1 and Line 3, it can be easily seen that the difficult parameter b shifts (or translates) the ICC along the x (or ability) axis. Line 3 represents a more difficult item because, given a fixed ability (say 2.5), the probability of a correct response is much lower in Line 3 than in Line 1. From the comparison between Line 1 and Line 2, it can be also easily seen that the discrimination parameter a is essentially a regression coefficient representing the strength (or quality) of the relationship. Around an ability of zero, Line 2 has more discrimination power than does Line 1 in that Line 2 shows greater differences in the predicted probability.

In the equation for the 2PL model, the distance between θ_j and b_i determines the probability of correct responses. For example, an easy item administered to a person with a high ability (i.e., $\theta_j - b_i >>$) produces a high probability of getting the item correct. For more details about IRT, readers are referred to the specialized literature [17,19].

In IRT, each student is assumed to have an underlying ability in a specific domain. The goal of IRT calibration (or estimation) is to locate each person's ability on a continuous latent variable. The strength of IRT mainly comes from the fact that IRT also locates the difficulty of each item on the same latent variable on which each person's ability is located. That is, IRT locates both the persons' abilities and the items' difficulties on the same latent continuum and decides relative distances among the persons' abilities and the items' difficulties on the continuum. In locating persons and items on a latent continuum (or estimating persons' and items' parameters), each IRT estimation (or calibration) needs to define its own metric because latent variables representing students' abilities do not have any inherent metric. For example, in IRT, the metric of a latent variable can be defined by fixing the variance of the latent variable to one. Figure 2 illustrates two separate IRT

calibrations having their own metrics. In the first calibration (IRT Calibration 1), three items (Item 1, Item 2, and Item 3) and two persons (Person 1 and Person 2) are located on the latent continuum. A larger value in the continuum represents higher-ability persons or harder items. For example, Person 2 has a higher ability than Person 1, and Item 3 is more difficult than Item 2. The second calibration (IRT Calibration 2) locates different sets of items (Item 2, Item 3, and Item 4) and persons (Person 3 and Person 4) on its own metric. In the figure, different locations of and distances between Item 2 and Item 3 in the two metrics indicate that the two metrics have their own origins and units.

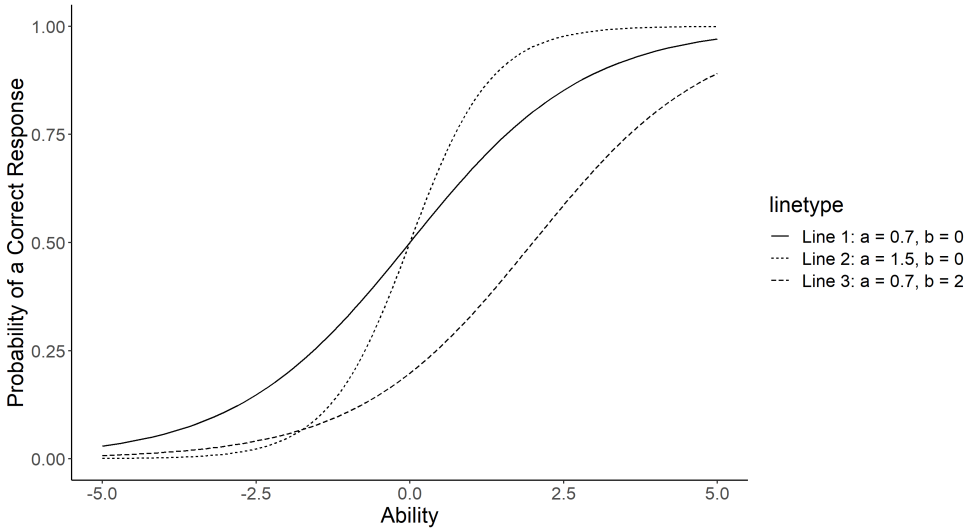

Figure 1. Item characteristic curves (ICCs) for three different sets of item parameters.

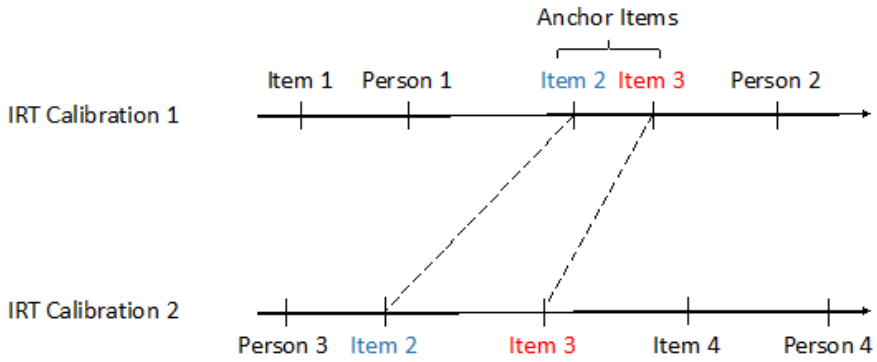

Figure 2. Illustration of two separate IRT calibrations having their own metrics.

The advantage of locating both the persons' abilities and the items' difficulties on the same latent continuum becomes clearer by contrasting IRT and classical test theory (CTT) [20]. In CTT, raw total scores are used to represent persons' abilities. Easier items would produce higher total scores, and harder items would produce lower total scores, when the same abilities are assumed. That is, in the total scores, the persons' abilities are confounded with the items' difficulties. Because of this confounding, it is difficult to compare persons' abilities from different tests in CTT. On the contrary, IRT estimates person and item parameters separately. Therefore, the estimated person parameters (or persons' abilities) are not confounded with the item difficulties. This separate estimation of person and item parameters in IRT enables us to compare persons' abilities on different metrics from different tests. For example, in Figure 2, we can compare the ability of Person 2 in IRT Calibration 1 and the ability of Person 3 in IRT Calibration 2 based on their relative locations to the common items, Item 2 and Item 3, in both metrics: Person 2 has

a higher ability than Person 3. The common items in two separate calibrations are often called anchor items connecting two metrics. In sum, the separate calibration of persons' abilities and items' difficulties in IRT provides desirable features: item parameters are not sample-dependent (item parameter invariance); scores describing persons' ability are not test-dependent (person parameter invariance).

2.3. Fixed-Parameter Calibration

In educational measurement, IRT has been widely used in placing persons' abilities from different calibrations on a common metric or in equating [21]. In practice, the equating allows us to compare persons' abilities from different test forms. For example, graduate record examinations (GREs) use multiple test forms for various reasons including item security, and equating allows us to place examinees' abilities from different forms on a common metric for comparison. This equating procedure is essential in comparing the performance of persons over repeated measures. Three common approaches in IRT have been used to capture academic growth in the literature: the linear transformation of separate calibrations, concurrent calibration, and fixed-parameter calibration (FPC). In the specific context of MOOCs, Meyer and Zhu [7] proposed to use equating for fair and equitable measurement of student learning in MOOCs. In their work, they used the linear transformation of separate calibrations. In this study, we propose to use FPC to establish a common metric for the measurement of learning in MOOCs because FPC allows us to establish a common metric more easily without the process of score transformation, or the FPC method establishes a common metric in a way that fits well in MOOCs settings.

The key question of equating in education measurement is how to place the new and old items on a common scale for comparison. Two different ways are possible for equating: independent calibration with linking and FPC [8]. In the independent calibration, we first calibrate (or estimate item parameters) the new and old item together and then place the calibrated parameters for the new items on the existing scale using the linear transformation that can be found by linking items. On the contrary, FPC does not require linking items, which makes FPC more suitable for the measurement of learning in MOOCs.

When calibrating IRT using FPC, the item parameters of a subset of items, called operational item parameters, are fixed at their previously calibrated values, and only the item parameters of new items are calibrated. As a result, the new items are placed on the same metric as the previously calibrated items [8,22]. Figure 3 illustrates FPC. In the figure, the item parameters of Item 2 and Item 3 in IRT Calibration 2 are fixed to their previously calibrated values in IRT Calibration 1, and only the parameters of Person 3, Item 4, and Item 5 are estimated in IRT Calibration 2. The fact that the item parameters of Item 2 and Item 3 are exactly the same in the two calibrations indicates that the two calibrations are on a common metric. FPC is simpler than separate calibrations because it does not require any transformation, but simply requires fixing some of the item parameters to their previously calibrated values during calibration. Furthermore, FPC is well suited for establishing a common metric across multiple administrations of a MOOC because those multiple administrations are highly likely to share a large number of items, which can be used as operational items in FPC. In educational measurement, FPC has been widely used in online calibration used in computerized adaptive testing (CAT) to calibrate new items in the tests on the same metric with previously calibrated items [23–25].

In addition to the common metric, FPC provides an additional advantage in the measurement of learning in MOOCs. Because of the nonmandatory and self-paced nature of MOOCs, typical MOOCs data contain a large number of missing responses to items [26]. In general, IRT is known to be robust to missing data [27]. As a result, IRT scores are expected to be more robust to missing data than the percentage or raw total scores.

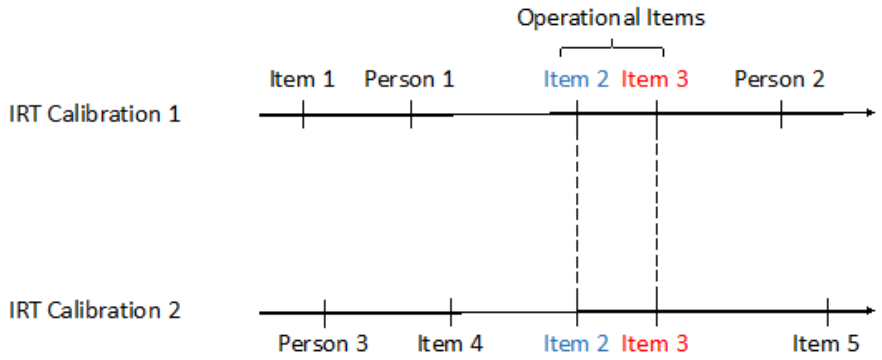

Figure 3. Illustration of fixed-parameter calibration (FPC).

3. Current Study

The measurement of learning is an essential task in MOOCs research. In this study, we propose to use FPC to establish a common metric for the measurement of learning in MOOCs. Our analysis was guided by the following research questions:

RQ1: Does a common metric make meaningful changes in students' abilities?
RQ2: Does a common metric make meaningful changes in students' gain scores?
RQ3: Are the IRT scores from FPC more robust to missing data than percentage scores?

In percentage scores, students' abilities and items' difficulties are confounded. A common metric is expected to disentangle the two confounding factors and, therefore, is expected to make meaningful changes in students' abilities and also gain scores. From the reasons described in the previous section, it was also expected that the IRT scores are more robust to missing data.

4. Methods

Samples for our analysis came from the MOOC "Introduction to Psychology as a Science" offered through Coursera collaboratively by Georgia Institute of Technology (GIT) and Carnegie Mellon University (CMU) in 2013. This course was designed as a typical 12-week introductory course for the first-year college students. A large number of interactive activities from the "Introduction to Psychology" course offered through the CMU's Open Learning Initiative (OLI) were embedded in this course in the Coursera platform using Learning Tools Interoperability (LTI). Topics covered in this course include memory, sense and perception, abnormal behavior, brain structures and the nervous system, and so on. For research purposes, this course contained pre- and post-tests. Students took the pre-test consisting of 20 multiple-choice items before the course started. At the end of the course, students took the final exam (post-test) consisting of 35 multiple-choice items. A more detailed description of this course can be found elsewhere [28,29]. Among the total of 27,720 students who registered the course, the data for our analysis consisted of 829 students (3%) who completed both the pre- and post-tests.

5. Data Analysis

FPC requires previously calibrated item parameters to fix item parameters for some items (or operational items). To mimic the situation in which the same MOOC is administered multiple times, we randomly split the 829 samples into two subsets: one (D1: $N = 414$) representing data from the past year and the other (D2: $N = 415$) representing data from the current year. Then, using D1, the total of 55 items (20 pre-test items and 35 post-test items) were calibrated using the 2PL model. This calibration located the difficulty parameters of all 55 items on a common metric. The calibration was conducted using the tam.mml.2pl() function in the TAM package in R [30]. In our study, the item parameters calibrated using D1 were considered as item parameters for the operational items in FPC. Then, using D2, the persons' abilities for pre- and post-tests were calibrated by fixing the item parameters

of all 55 items at their previously calibrated values. In general, the operational items in FPC are some subset of items. In our study, we split a given dataset into two subsets to mimic the situation in which the same MOOC is administered multiple times, and therefore, all items were considered as the operational items. Finally, the persons' abilities for both the pre- and post-tests from FPC were considered as the persons' abilities on a common metric. RQ1 and RQ2 were addressed by comparing the persons' abilities from FPC and raw percentage scores. To address RQ3, artificial random missing with different missing proportions (50% and 70%) was introduced to the data. Then, the robustness of the IRT scores from FPC and raw percentage scores was evaluated.

6. Results

6.1. Percentage Scores

Figure 4 presents the percentages of correct responses to the items and the persons in the pre- (colored blue) and post-tests (colored orange) using box plots. The plots clearly show that the percentages of correct responses were higher in the post-test. For the measurement of students' learning, the key question was whether students' higher proportion of correct answers on the post-test should be attributed to students' actual learning or to easier items in the post-test. The two factors (i.e., students' abilities and items' difficulties) were confounded in percentage scores. Therefore, percentage scores cannot be used to answer this important question.

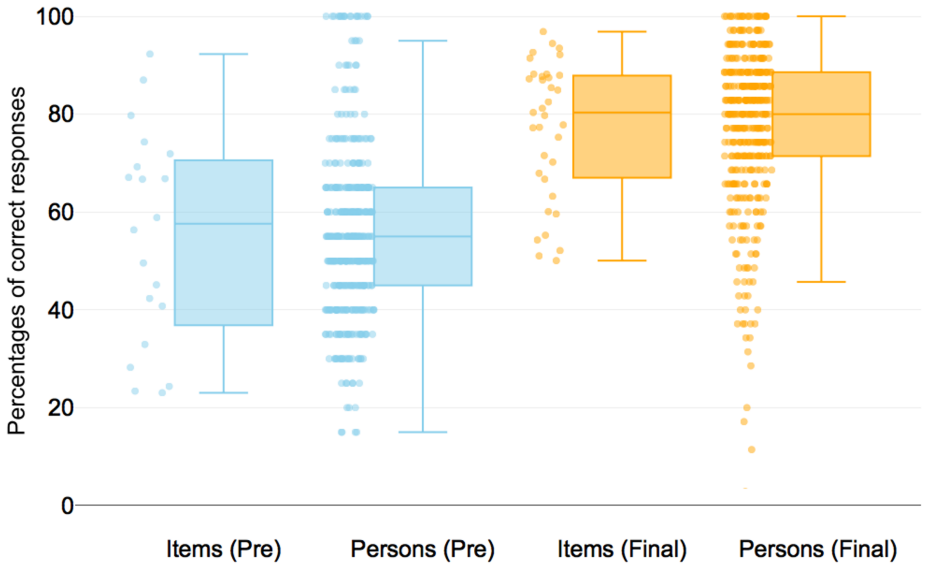

Figure 4. Percentages of correct responses to items and persons in the pre- and post-tests.

6.2. Percentage Scores vs. IRT Scores

Figure 5 presents the comparison between the percentage scores and IRT scores. The IRT scores for both the pre- and post-tests are located on a common metric because those scores come from FPC. Figure 5a presents the scatter plot between percentage scores (y-axis) and IRT scores (x-axis) for both the pre- (colored blue) and post-tests (colored orange). This plot clearly shows that the same percentage score can be further differentiated using the IRT scores. For example, the percentage scores of around 60 corresponded to the IRT scores of around -1 in the post-test (colored orange), whereas the same percentage scores of around 60 corresponded to the IRT scores of around 0 in the pre-test (colored blue). Because the IRT scores from FPC were located on a common metric, the differences in the IRT scores between the pre- and post-tests can be attributed to the students' actual learning without worrying about the confounding between the students' abilities and the items' difficulties.

Figure 5b presents the scatter plot between the two gain scores based on the percentage scores (y-axis) and IRT scores (x-axis). There was a strong correlation between the two gain scores (r = 0.96), indicating that the order structure of the gain scores was generally preserved. However, the two gain scores produced quite different classification results in terms of students' learning. In Figure 5b, the green dots represent a students who had positive learning gains based on the percentage scores, but had negative learning gains based on the IRT scores. Table 1 presents the cross-tabulation between the two gain scores. There were 349 students who had positive learning gains based on the percentage gain scores, whereas 206 students had positive learning gains based on the IRT gain scores. There were 140 students who had positive learning gains based on the percentage scores, but had negative learning gains based on the IRT scores. The classification based on the IRT scores should be more reliable because the IRT scores for the pre- and post-tests were placed on a common metric and, therefore, were free from the confounding between the students' abilities and the items' difficulties.

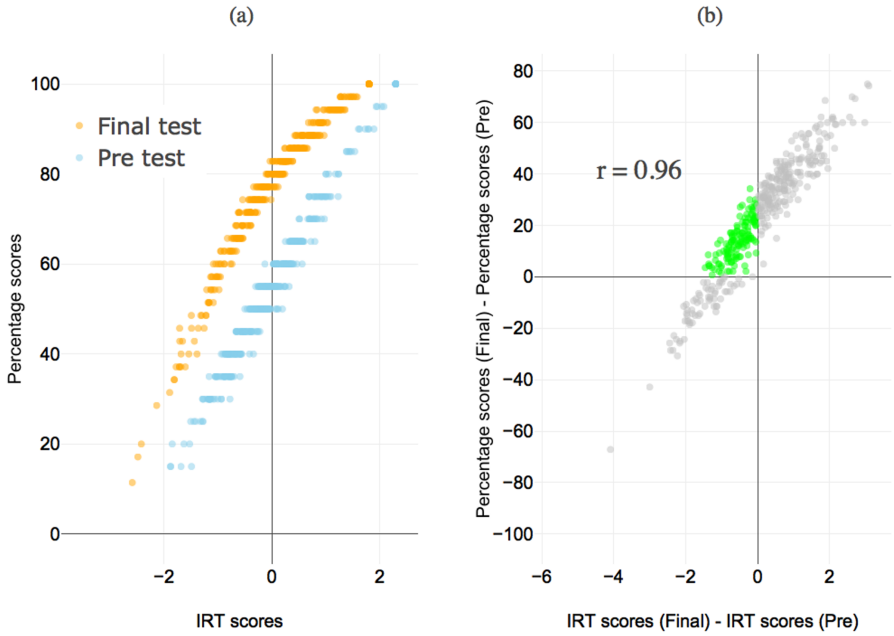

Figure 5. Comparison between the percentage and IRT scores (**a**,**b**).

Table 1. Gain scores based on the percentage and IRT scores.

		Final-Pre		
		Negative	**Positive**	**Total**
Final-Pre	Positive	140	209	349
	Negative	66	0	66
	Total	206	209	415

Figure 6 illustrates the percentage and IRT scores between the pre- and post-tests using box plots. By visual inspection, the percentage scores seemed to show significant learning gains, whereas the IRT scores seemed not to show significant learning gains. This observation was confirmed by the repeated-measures t-tests presented in Table 2, which suggested that the learning gains based on the percentage scores might be spurious.

Table 2. Repeated-measures *t*-tests.

	M(SD) for Pre	M(SD) for Post	Mean Difference	t	*p*-Value
Percentage scores	55.204 (17.819)	77.562 (16.620)	22.358	20.325	<0.001
IRT scores	0.079 (0.803)	0.093 (0.942)	0.013	0.246	0.805

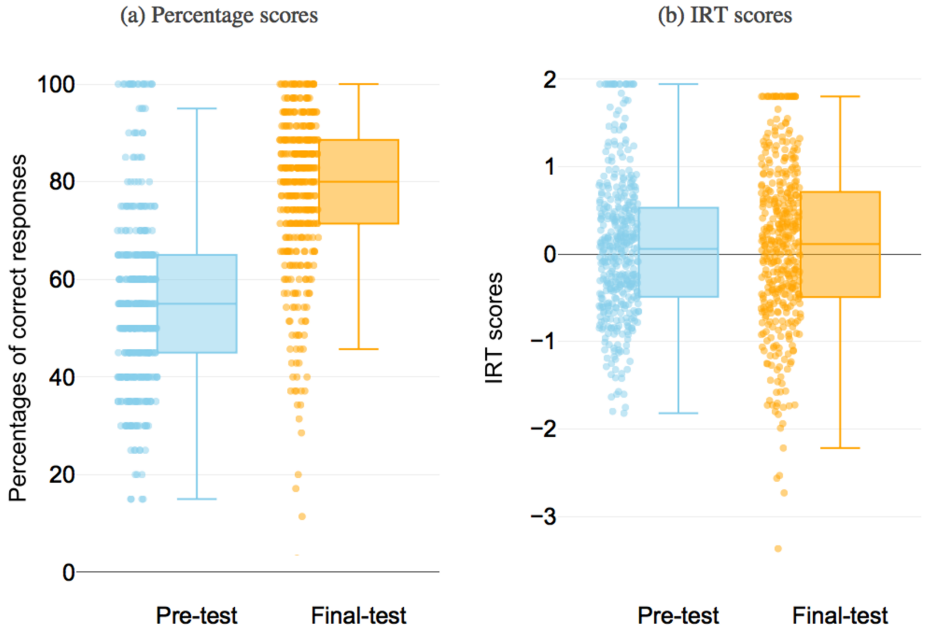

Figure 6. Learning gains between the pre- and post-tests.

7. Discussion

The measurement of learning is an essential task in E-learning. For the valid measurement of learning in E-learning, scores must be placed on a common metric or equated. Despite the crucial role of a common metric, however, less attention has been paid to establishing a common metric in the measurement of learning in E-learning. This study aimed to fill this gap in the E-learning literature. Among various equating methods that have been used in the field of educational measurement, we proposed to use FPC, which establishes a common metric in a way that fits well in E-learning settings. A common metric was expected to allow us (a) to define students' learning properly, (b) to evaluate students fairly, (c) to reuse E-learning items efficiently, and (d) to prevent cheating in E-learning. We used the data from the MOOC "Introduction to Psychology as a Science" offered through Coursera collaboratively by Georgia Institute of Technology (GIT) and Carnegie Mellon University (CMU) in 2013 to answer our specific research questions: Does a common metric make meaningful changes in students' abilities and gain scores? Are the IRT scores more robust to missing data than percentage scores? The results from our analysis showed that a common metric made substantial changes in the students' abilities and gain scores. These results suggested that IRT scores are preferable in E-learning.

As presented in Figure 5a, the same percentage score in the pre- and post-tests was mapped into different IRT scores in the pre- and post-tests, which implies that the same percentage score (say 60% of correct responses) in the pre- and post-tests may represent different levels of students' abilities (RQ1). This makes sense because the pre- and post-tests included different sets of items, and therefore, the same percentage score in the pre- and post-tests may represent different levels of the students' abilities. The different IRT scores corresponding to the same percentage scores were the result of removing confounding between the students' abilities and the items' difficulties in the IRT models. In our analysis, the percentage scores tended to overestimate the students' learning gains

(RQ2). According to Table 1, 140 students (about 40%) who were classified as having positive learning gains based on percentage scores were reclassified as students having negative learning gains based on the IRT scores. The repeated-measures *t*-tests in Table 2 also demonstrated that the overall learning gains shown in the percentage scores (mean difference = 22.358, *p*-value = 0.000) disappeared when using the IRT scores (mean difference = 0.013, *p*-value = 0.805).

One of the limitations of our study was the splitting of our MOOC data to mimic the multiple administrations of a MOOC. FPC requires previously calibrated item parameters to fix some of the item parameters during the calibration. In real settings, those previously calibrated item parameters can be obtained from the previous administrations of a MOOC. However, in our study, our access to the data was restricted to the 2013 MOOC data. Therefore, we randomly split our data into two subsets representing data from the past and current year. A study using real longitudinal data would be an interesting future work.

Assessment is the key to identify factors that can promote students' learning in E-learning. For valid assessment, scores must be placed on a common metric or equated. This study contributes to the E-learning literature by highlighting the importance of a common metric and also by introducing FPC as a simple equating procedure.

Author Contributions: Conceptualization, S.L., Y.-J.C. and H.-S.K.; methodology, S.L.; software, S.L.; validation, S.L., Y.-J.C. and H.-S.K.; formal analysis, S.L.; investigation, S.L., Y.-J.C. and H.-S.K.; resources, S.L.; data curation, S.L.; writing—original draft preparation, S.L.; writing—review and editing, S.L., Y.-J.C. and H.-S.K.; visualization, S.L.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ewha Womans University Research Grant of 2020. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5C2A03093092).

Institutional Review Board Statement: Ethical review and approval were waived for this study because this was a secondary data analysis.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: https://pslcdatashop.web.cmu.edu/, accessed on 1 September 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kennedy, J. Characteristics of massive open online courses (MOOCs): A research review, 2009–2012. *J. Interact. Online Learn.* **2014**, *13*, 1–16.
2. Zhu, M.; Sari, A.; Lee, M.M. A systematic review of research methods and topics of the empirical MOOC literature (2014–2016). *Internet High. Educ.* **2018**, *37*, 31–39. [CrossRef]
3. Reich, J. Rebooting MOOC Research. *Science* **2015**, *347*, 34–35. [CrossRef] [PubMed]
4. de Barba, P.; Kennedy, G.; Ainley, M. The role of students' motivation and participation in predicting performance in a MOOC. *J. Comput. Assist. Learn.* **2016**, *32*, 218–231. [CrossRef]
5. Kulkarni, C.E.; Bernstein, M.S.; Klemmer, S.R. PeerStudio. In Proceedings of the Second (2015) ACM Conference on Learning @ Scale, Vancouver, BC, Canada, 14–18 March 2015.
6. Marin, V.J.; Pereira, T.; Sridharan, S.; Rivero, C.R. Automated Personalized Feedback in Introductory Java Programming MOOCs. In Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 19–22 April 2017.
7. Meyer, J.P.; Zhu, S. Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Res. Pract. Assess.* **2013**, *8*, 26–39.
8. Kim, S. A Comparative Study of IRT Fixed Parameter Calibration Methods. *J. Educ. Meas.* **2006**, *43*, 355–381. [CrossRef]
9. Huggins, A.C.; Penfield, R.D. An NCME Instructional Module on Population Invariance in Linking and Equating. *Educ. Meas. Issues Pract.* **2012**, *31*, 27–40. [CrossRef]
10. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
11. Dugard, P.; Todman, J. Analysis of Pre-test-Post-test Control Group Designs in Educational Research. *Educ. Psychol.* **1995**, *15*, 181–198. [CrossRef]

12. Alario-Hoyos, C.; Estévez-Ayres, I.; Pérez-Sanagustín, M.; Kloos, C.D.; Fernández-Panadero, C. Understanding Learners' Motivation and Learning Strategies in MOOCs. *Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*. [CrossRef]
13. Seaton, D.; Hansen, J.; Goff, J.; Sellers, P.; Davidson, N.; Houck, A. Transforming Advanced Placement High School Classrooms Through Teacher-Led MOOC Models. Digital Inclusion: Transforming Education Through Technology. In Proceedings of the LINC (Learning International Networks Consortium), Cambridge, MA, USA, 23–25 May 2016; p. 237.
14. Rogers, C.F. Faculty perceptions about e-cheating during online testing. *J. Comput. Sci. Coll.* **2006**, *22*, 206–212.
15. Underwood, J.; Szabo, A. Academic offences and e-learning: Individual propensities in cheating. *Br. J. Educ. Technol.* **2003**, *34*, 467–477. [CrossRef]
16. Rowe, N.C. Cheating in online student assessment: Beyond plagiarism. *Online J. Distance Learn. Adm.* **2004**, *7*. Available online: https://www.westga.edu/~distance/ojdla/summer72/rowe72.html (accessed on 1 September 2021).
17. Embretson, S.E.; Reise, S.P. *Item Response Theory for Psychologists*; Psychology Press: New York, NY, USA, 2013.
18. Harris, D. Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educ. Meas. Issues Pract.* **1989**, *8*, 35–41. [CrossRef]
19. De Ayala, R.J. *The Theory and Practice of Item Response Theory*; Guilford Publications: New York, NY, USA, 2013.
20. Hambleton, R.K.; Jones, R.W. Comparison of classical test theory and item response theory and their applications to test development. *Educ. Meas. Issues Pract.* **1993**, *12*, 38–47. [CrossRef]
21. Cook, L.L.; Eignor, D.R. IRT Equating Methods. *Educ. Meas. Issues Pract.* **1991**, *10*, 37–45. [CrossRef]
22. Jodoin, M.G.; Keller, L.A.; Swaminathan, H. A Comparison of Linear, Fixed Common Item, and Concurrent Parameter Estimation Equating Procedures in Capturing Academic Growth. *J. Exp. Educ.* **2003**, *71*, 229–250. [CrossRef]
23. Ban, J.C.; Hanson, B.A.; Wang, T.; Yi, Q.; Harris, D.J. A Comparative Study of On-line Pretest Item-Calibration/Scaling Methods in Computerized Adaptive Testing. *J. Educ. Meas.* **2001**, *38*, 191–212. [CrossRef]
24. Stocking, M.L. Scale drift in on-line calibration. *ETS Res. Rep. Ser.* **1988**, *1988*, i-122.
25. Wainer, H.; Mislevy, R.J. Item response theory, item calibration, and proficiency estimation. *Comput. Adapt. Test. A Primer* **1990**, *4*, 65–102.
26. Bergner, Y.; Colvin, K.; Pritchard, D.E. Estimation of ability from homework items when there are missing and/or multiple attempts. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY, USA, 16–20 March 2015.
27. Mislevy, R.J.; Wu, P.K. Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Res. Rep. Ser.* **1996**, *1996*, i-36. [CrossRef]
28. Koedinger, K.R.; Kim, J.; Jia, J.Z.; McLaughlin, E.A.; Bier, N.L. Learning is Not a Spectator Sport. In Proceedings of the Second (2015) ACM Conference on Learning @ Scale, Vancouver, BC, Canada, 14–18 March 2015. [CrossRef]
29. Koedinger, K.R.; McLaughlin, E.A.; Jia, J.Z.; Bier, N.L. Is the doer effect a causal relationship? In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, UK, 25–29 April 2016. [CrossRef]
30. Robitzsch, A.; Kiefer, T.; Wu, M. *TAM: Test Analysis Modules*; R Package Version 3.7-16; The R Foundation: Vienna, Austria, 2021.