THE EQUIVALENCE OF TOEP FORMS

Suwarsih Madya[a]
Heri Retnawati[b], Ari Purnawan[c],
Nur Hidayanto Pancoro Setyo Putro[d], Ezi Apino[e]

([a]suwarsihmadya@uny.ac.id; [b]heri_retnawati@uny.ac.id;
[c]ari_purnawan@uny.ac.id; [d]nur_hidayanto@uny.ac.id;
[e]apinoezi@gmail.com)

Universitas Negeri Yogyakarta
Colombo Street No 1, Depok, Sleman, Yogyakarta, Indonesia

Abstract: This explorative-descriptive study set out to examine the equivalence among Test of English Proficiency (TOEP) forms, developed by the Indonesian Testing Service Centre (ITSC) and co-founded by The Association for The Teaching of English as a Foreign Language in Indonesia (TEFLIN) and The Association of Psychology in Indonesia. Using a quantitative approach, the researchers collected the data through documenting the responses of those taking TOEP in 2016 and 2017, involving six TOEP forms in 2016 and four TOEP forms in 2017. All the forms were developed using the same test grid and construct to measure the listening and reading skills. The equality among the six forms was tested using the equating technique, which involved (1) the estimation of the item parameter using the Rasch model, (2) examination of the test characteristics curve for each form, and (3) interpretation of results. The results showed that all the TOEP forms used in 2016 and 2017 were equal with one another. It can be concluded then that the developed TOEP forms have the same level of difficulty and ensure justice for all test takers.

Keywords: equating, test forms, TOEP, Rasch model, test of English proficiency

DOI: http://dx.doi.org/10.15639/teflinjournal.v30i1/88-104

In this globalized and globalizing digital world, people from different countries interact with one another, either face-to-face or virtually for different purposes.

This interaction occurs without any space and time constraints due to the advancements of communication and information technology (West, 2010). Central to this borderless interaction is the role of international languages in bridging the information gaps. With their status as official languages in the United Nations, six languages are recognized as international languages, i.e. Arabic, Chinese, English, French, Russian, and Spanish. However, in terms of users and the scope of use, English seems to be the most widely used language in the world as mentioned in Sawe (2019). Mastering English can therefore be said to be a fundamental need for those who want to succeed in this globalized and globalizing digital world.

In Indonesia, English is the first and only foreign language taught as a compulsory subject at all levels of education and learned in thousands of private courses across the country. All courses, be they in schools or private institutions, aim to develop the participants' English language skills –Listening, Speaking, Reading, and Writing. Different courses might be designed differently, but the results of learning may be measured through a standardized test so that the test scores can be recognized across institutions, regions and countries. Despite the fact that IELTS and TOEFL are today's most known standardized English tests, with a big population of English learners in Indonesia and the fact that English is a foreign language, the development of a standardized test of English in this country is considered a necessity. The standardized English test can be used to measure the results of English learning regardless of what courses are run, in what places they are held, and when they are run.

Considering the needs for measuring the results of English learning as touched upon before, the Association for the Teaching of English as a Foreign Language in Indonesia (TEFLIN) and the Association for Psychology in Indonesia have jointly founded the Indonesian Testing Service Centre (ITSC). This centre develops an online-based TOEP (Test of English Proficiency) and a Basic Academic Potential Test. These two tests have been taken by more than 100 thousand academicians and postgraduate students in more than 50 test centers all over the country. The results of TOEP provide evidence of the success of the English courses and learning in general. This means that the test scores can be used as input in evaluating the English courses taken by the test takers. As stated in Brookhart and Nitko (2015), testing is one of the assessment activities that can be used to evaluate the significance of a program, education intervention, a curriculum model, a pedagogical initiative, or a policy

in the language field of study (Brown, 2004; Gitsaki & Robby, 2018). In addition, from the point of view of learning, especially learning English in private courses, an assessment is necessary to provide teachers/instructors with evidence of how successful their instruction has been and to provide the basis for an evaluation of the program (Linn & Gronlund, 2005).

In general, an English test measures the test takers' proficiency of listening, speaking, reading and writing (Foster, 2009; Rahman, Babu, & Ashrafuzzaman, 2011). In reference to Norris (2000), TOEP is developed and administered as a procedure or an instrument used for collecting information on test takers' English proficiency. For practical reasons, TOEP is temporarily limited to the English receptive skills, i.e. listening and reading.

For all adult learners of English with different needs for English proficiency in this digital era, measuring the learners' English proficiency through an online standardized test has a lot of benefits (Bartram, 2008). One of the benefits is that a greater number of people can take the test without time and space constraints. With a vast area in the Indonesian archipelago, in which transportation is not yet an easy solution for space constraints, an online test may be much more efficient in terms of resources than a paper-and-pencil test. Another benefit is that it is easier to manage the test, especially in high-stakes situations. In terms of the scoring and reporting, online tests allow easier and faster scoring and reporting compared to paper and pencil tests.

Apart from the benefits, there are some shortcomings of an online test. The first one is its dependency on the availability of necessary equipment and reliable technicians. To overcome this shortcoming, a standard operational procedure has been set for TOEP administration. This is to ensure that the test can securely run well. Another shortcoming is related to security, in a way that test takers might take the test several times until they remember all the questions. There is also a possibility that the different test forms are not equal; hence, possible injustice for test takers. Test administrators and developers have made efforts to overcome the shortcomings by creating different forms of tests and ensuring that the forms are equivalent (Baghaei, 2010; Kartowagiran, Munadi, Retnawati, & Apino, 2018). This article focuses on examining the equivalence among forms of TOEP that have been used widely in Indonesia.

Ensuring the equivalence of test forms is necessary and this can be conducted through the equating technique, which generally aims to place the score of two or more tests on the same scale (Hambleton, Swaminathan, & Rogers, 1991), to find out if the two or more tests are equivalent or not

(Kartowagiran et al., 2018). The equating technique is part of the measurement science, which has been used in different countries.

The equating of tests can be conducted by using both classical and modern approaches or the so-called the item-response theory (Ryan & Brockmann, 2009). In the modern approach the equating is conducted by calculating item parameters (difficulty index, discrimination index, and pseudo-guessing index) and the test takers' ability parameters toward a score using a linier equation (Retnawati, 2016). Before conducting the calculation, an estimation toward the item parameters and ability parameters must be made first. If the item parameter used in making the estimation is limited to only the difficulty index (b), it is called estimation of 1 Parameter Logistic (1 PL). If the estimation is conducted by using the difficulty index (b) and the differentiating power index (a), it is called estimation of 2 Parameter Logistic (2 PL). If the item parameter covers three parameters altogether, i.e. the difficulty index (b), the differentiating power index (a), and pseudo-guessing index, it is called estimation of 3 Parameter Logistic (3 PL) (Embretson & Reise, 2013).

After the estimation of each parameter has been obtained, the subsequent step is to find out the inter-test form equivalence by using the equating method. In this case, there are different methods which can be used: the means-means method, the means-sigma method, and the curve of item characteristic which include the Haebara method and the Stocking and Lord methods (Hambleton et al., 1991; Kolen & Brennan, 2004). The means-means method and the means-sigma method need special attention because these two methods apply a simple equation and its application is also very easy (Retnawati, 2014). In addition, according to Hambleton et al. (1991), in the means-means method there is a reciprocal relationship, meaning that if tests X and Y are correlated, the correlation between Y and X can be determined.

The linear equation formed in the equating technique employs a constancy, i.e. α and β. In the means-means method, the equivalence constancy α and β can be calculated by using the means from the means from the item difficulty index (b) and that of the item differentiating power (a) (Hambleton et al., 1991). For example, if test X will be equated to test Y by using the 3 PL model, the relationship between the difficulty index parameter (b) and the differentiating power parameter (a) can be formulated as $b_y = \alpha b_x + \beta$ and $a_y = \frac{a_x}{\alpha}$, with b_x dan b_y being the difficulty index of test X and test Y, a_x and a_y as the differentiating power of tests and Y. Because the constancy is

calculated by using the difficulty index means (*b*) and the differentiating power index means (*a*), it will result in $\bar{b}_y = \alpha\bar{b}_x + \beta$ so that $\beta = \bar{b}_y - \alpha\bar{b}_x$ and $\bar{a}_y = \frac{\bar{a}_x}{\alpha}$ so that $\alpha = \frac{\bar{a}_x}{\bar{a}_y}$ with $\bar{b}_x$ dan $\bar{b}_y$ being the difficulty index means of tests X and Y with $\bar{a}_x$ and $\bar{a}_y$ being the differentiating power between tests X and Y.

In the means-sigma method, the calculation to the constancy equating α and β is conducted by involving the means and standard deviation of the parameter of the item difficulty index (Hambleton et al., 1991). For example, if test X will be equated to test Y by using the 3 PL model, the equating model can be formulated as $b_y = \alpha b_x + \beta$ and $S_y = \alpha S_x$, with S_x and S_y being the standard deviation of the item difficulty index of tests X and Y. Since the constancy is calculated by using the means and standard deviation of the difficulty index, it will result in $\bar{b}_y = \alpha\bar{b}_x + \beta$ so that $\beta = \bar{b}_y - \alpha\bar{b}_x$ and $\alpha = \frac{S_y}{S_x}$, with $\bar{b}_x$ dan $\bar{b}_y$ being the means of the difficulty indices of tests X and Y with S_x and S_y being the standard deviation of the difficulty index of tests X and Y.

The next example is if the scale test X I equated to the scale of test Y for the 3 PL model, the relationship between item parameters (a, b, dan c), according to Kolen and Brennan (2004), the parameter of test taker ability (θ) and the equating constancy (α and β) for the two scales can be formulated as $\theta'_{xi} = \alpha\theta_{xi} + \beta$, with $a'_{xj} = \frac{a_{xj}}{\alpha}$, $b'_{xj} = \alpha b_{xj} + \beta$, and $c'_{xj} = c_{xj}$, where θ_{xi} is the test takers' ability toward-i at the scale of test X; θ'_{xi} is the test takers' ability toward-i at the scale of test X after being equated with test Y; α_{xj}, b_{xj}, and c_{xj} is the item parameter toward-j at the scale of test X; and a'_{xj}, b'_{xj}, and c'_{xj} are the item parameter for item toward -j at the scale of test X after being equated with test Y. It should be noted that the item parameter c is not transformed because the value of parameter *c* does not depend on the ability parameter (θ), so that parameter c is free from the scale transformation (Kolen & Brennan, 2004). This means that the value of c at the scale of test X will remain the same as the value of c at the scale of test X which has been equated with test Y.

Other than the methods already mentioned, there are a lot of equating methods which can be used. It should be noted, however, that each method has its own strengths and weaknesses. In relation to this, the finding of a research

study conducted by Pang, Madera, Radwan, and Zhang (2010) indicates that the Stocking and Lord method of equating the curve of characteristics and the means-sigma method give the equivalent good result. The finding of another research study by Retnawati (2016) indicates that graphically, the means-sigma method gives the most equated score compared to the Haebara's and Stocking and Lord's means-means method. Meanwhile, Yu and Popp (2005) in their research report stated that there was no one best method for equating test scores. Smith and Kramer (1992) suggest that the selection of equating techniques depends on the test characteristics and sample used in equating. The researcher can therefore select one method of equating to examine the equivalence between test forms.

Some research studies which aimed to provide evidence of test forms equivalence have been conducted. One of them was by Yim and Huh (2006), who investigated the equivalence of the Medical Licensing Examination forms in 2003 and 2004 by employing the item response theory approach, the results of which indicated that the Medical Licensing Examination in 2003 was more difficult than that in 2004. Sutrisno's (2016) study which investigated the quality of the school mathematics test in Bangkalan Regency, Indonesia, found that by using the test characteristics curve, the five test forms examined were found to be well equated. A study by Kartowagiran et al. (2018) investigated the equivalence of the National Examination forms in Indonesia from 2013-2016 by using the means-sigma. The equating technique proved that the test forms tended to be equitable.

Based on the background and the literature review and the results of the previous studies, proving the equivalence of test forms is a fundamental step to find out the quality of the test forms already developed. In this way, the objective of this study is to examine the equivalence of TOEP forms already developed by TEFLIN in the ITSC. The result of this study is expected to provide inputs to the test developers to produce standardized test forms so that a follow-up action can be taken to continue improving the construction of TOEP online.

METHOD

This research study was an explorative-descriptive one using quantitative approach to examine the equivalency of the TOEP farms developed by the ITSC (Indonesian Testing Service Centre) with TEFLIN holding the substantial

responsibility. The TOEP forms investigated in this study are those developed and used in 2016 and those in 2017. These two forms were selected since the other forms of tests had been equated before. Each test form measures the test takers' listening and reading proficiency. The 2016 TOEP consists of six forms, while the 2017 four forms. Each test form consists of 50 multiple choice listening test items and 50 multiple choice reading test items, with each item having one answer key. All the forms have been developed based on the same grid and construct.

Data collection was conducted through documenting test takers' responses in 2016 (N= 4448) and 2017 (N= 707). Data analysis was conducted by using the item response theory to examine the equivalency of TOEP forms. In this case, the Rasch model was used to analyze item characteristics with the item difficulty index being the only item parameter being measured. The estimation of the item parameter was conducted by using the QUEST software. After obtaining the value of the item parameter of each test form, the next thing to do was to make the test characteristics curve of each form. The test characteristics curve was made in reference to the mathematical model 1 of logistic equation (the Rasch model). Each test characteristics curve was then presented in the same field and scale. When the curves were overlapping one another, the test forms were proved to be equivalent. The more the test forms overlap, the more equivalent the test forms with one another will be (Smith & Kramer, 1992).

FINDINGS AND DISCUSSION

Findings

The Item Characteristics of the 2016 and 2017 TOEP Forms

This study examined the 2016 and 2017 TOEP forms, each of which consists of the listening and reading proficiency test sections. The 2016 TOEP forms analyzed consist of six test forms, while the 2017 TOEP of four forms. The related statistics of the test form characteristics are first presented before the test characteristics curve showing the equivalency of the TOEP test forms. The test form characteristics were analyzed using the Rasch model, so that the item parameter being estimated was only the item difficulty index (*b*). The estimation statistics of the item difficulty index of the 2016 TOEP and 2017 TOEP are presented in Tables 1 and 2.

Table 1. The Estimation Statistics of the Item Difficulty Index of the Listening Test

Statistics	The 2016 TOEP Forms						The 2017 TOEP Forms			
	1	2	3	4	5	6	1	2	3	4
Mean	0.000	0.000	-0.001	0.001	-0.001	0.000	0.000	0.000	0.000	0.000
Min	-1.950	-1.950	-2.910	-1.760	-1.940	-1.820	-2.080	-2.200	-2.100	-1.970
Max	1.390	1.390	1.310	1.480	1.380	1.340	2.380	1.870	3.250	2.260
SD	0.686	0.686	0.844	0.680	0.741	0.685	1.070	0.875	1.210	0.880

Theoretically, a test has a good quality if it has an item difficulty index between -2 and 2. Table 1 indicates that the type of listening test in the 2016 TOEP forms and 2017 TOEP forms has almost the same means of difficulty indices, i.e. at 0,000. This indicates that in general the 2016 and 2017 test forms have the difficulty index in the good category. If the minimum value of the item difficulty index is examined more carefully, some test forms have a minimum value of <-2, i.e. Form 3 of the 2016 TOEP, Form 1 of the 2017 TOEP, Form 2 of the 2017 TOEP, and Form 3 of the 2017 TOEP. This indicates that the test forms have the item difficulty indices in the unsatisfactory category, in which the difficulty index of < -2 indicates that the items are too easy. Meanwhile, concerning the maximum value of the test forms, Table 1 shows that some test forms have the difficulty indices of > 2, i.e. Form 1 of the 2017 TOEP, Form 3 of the 2017 TOEP, and Form 4 of the 2017 TOEP. This indicates that the listening test forms have items with a too high difficulty index. Then, if the standard deviation (SD) is examined, Table 1 shows that in the 2016 TOEP listening test, the distribution of the difficulty indices of Forms 3 and 4 distribute more widely from the means if compared to the other test forms. Meanwhile, for the 2017 TOEP listening, the difficulty indices of Form 1 and Form 3 also distribute more widely from the means if compared to the other test forms.

Table 2 in the next page presents the statistical data on the results of the estimation of the difficulty indices of reading test at the 2016 TOEP and 2017 TOEP forms. In Table 2 it can be seen that the means of the difficulty indices of each form is around 0,000. This indicates that in general the reading test forms in the 2016 TOEP and 2017 TOEP have the difficulty indices in the

good category. However, if the minimum value of the difficulty indices is carefully examined, it can be seen that in the 2016 TOEP there were five forms with the value of < -2, i.e. Form 1, Form 2, Form 3, Form 5 and Form 6. This indicates that in the forms some items are too easy. Meanwhile, for the 2017 TOEP, only one form with the difficulty index of <-2, i.e. Form 1. Then, if the maximum value of the difficulty indices as presented in Table 2 is examined carefully, there is one form in the 2017 TOEP with the difficulty index of >2, i.e. Form 3. This indicates that in this form there are items which are too difficult. Meanwhile, for the 2017 TOEP, three forms have items with high difficulty indices, i.e. Forms 1, 2 and 4. Then, if the standard deviation (SD) of the item difficulty indices is examined, the reading test forms in the 2016 TOEP with standard deviations which are not so different. Meanwhile, for the 2017 TOEP, the SD of each test form is more varied. This indicates that the distribution of the difficulty indices of the means of each form is not the same or more varied.

Table 2. The Statistics of the Estimation of the Difficulty Indices of the Reading test Items

Statistics	The 2016 TOEP Forms						The 2017 TOEP Forms			
	1	2	3	4	5	6	1	2	3	4
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	-0.001
Min	-2.330	-2.330	-2.680	-1.570	-2.460	-2.230	-3.010	-1.810	-1.860	-1.850
Max	1.460	1.460	2.590	1.680	1.850	1.570	3.720	3.420	1.710	2.050
SD	0.793	0.785	0.896	0.844	0.838	0.826	1.234	1.086	0.724	1.005

The Evidence of the Equivalence of the 2016 TOEP and the 2017 TOEP

As mentioned earlier, the 2016 TOEP consists of six test forms. The equivalence of the six test forms was examined by using the curves of the test characteristics. The closer the six curves of characteristics of the test forms are to one another, the more equivalent the test forms will be. For the listening test of the 2016 TOEP, the analysis of the equivalence of the six forms is presented in Figure 1, while that for the reading test in Figure 2.

Figure 1 shows that the test characteristics curve of each test form overlap one another, even perfectly for all the ability scales (-4 to +4). If examined in

more details, for all the three ability scales, i.e. low ((θ < -2), medium (-2 ≤ θ ≤ 2), and high (θ > 2), all the curves overlap. This indicates that based on the test takers' responses, the test forms are equivalent, for test takers of the three levels of ability–low, medium, and high. In this way, this indicated that the six listening test forms of the 2016 TOEP are equivalent with one another.

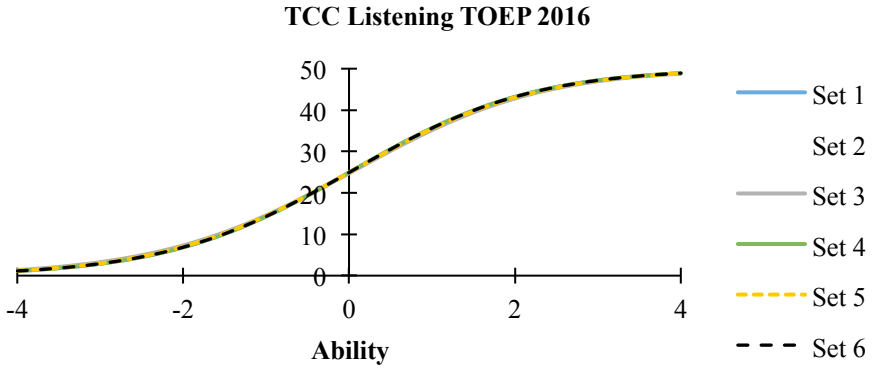

Figure 1. The Evidence of the Equivalence of the Six Listening Test Forms of the 2106 TOEP

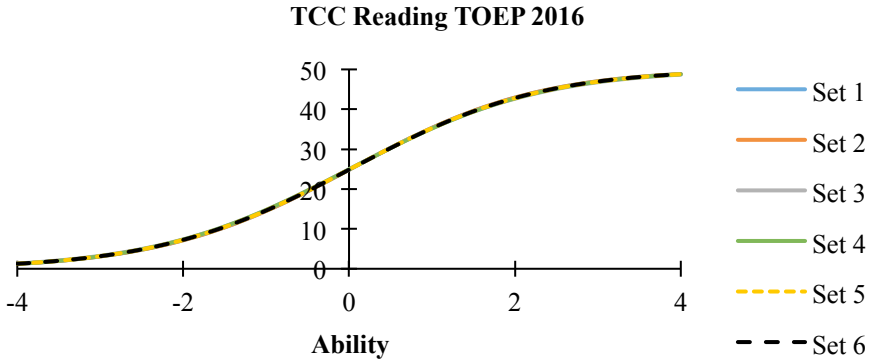

Figure 2. The Analysis of the Equivalence of the Six Reading Test Forms of the 2016 TOEP

The same case is shown in Figure 2 in the previous page, in which the characteristics curves of the test forms overlap one another for all ability scales of test takers. This indicates that the six reading test forms of the 2016 TOEP are equivalent with one another for test takers of all levels of ability –low, medium, and high. It can be interpreted that the difficulty index of the listening and reading test forms of the 2016 TOEP are relatively equivalent so that no test takers were disadvantaged.

The 2017 TOEP four test forms were investigated in terms of their equivalence. Similar to the 2016 TOEP, the equivalence of the four test forms was proved through examining the characteristics curves of the test forms. The analysis of the equivalence of the four listening and reading test forms are presented in Figure 3 and Figure 4.

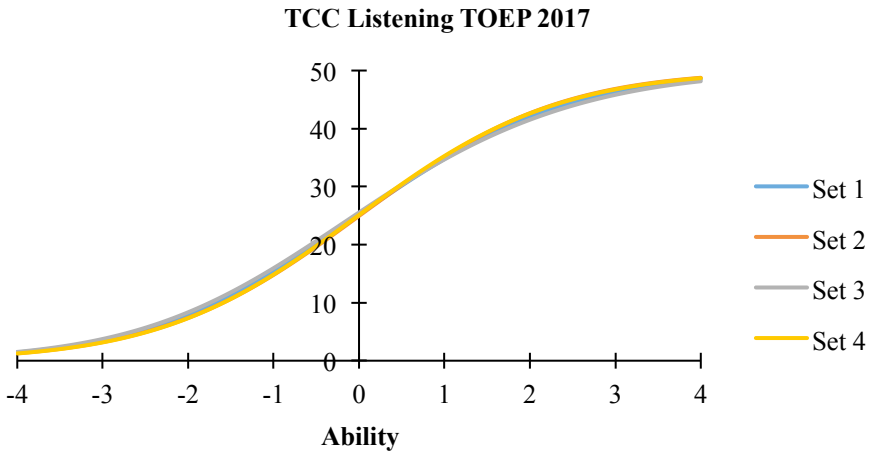

Figure 3. The Analysis of the Equivalence of the Four Listening Test Forms of the 2017 TOEP

Figure 3 shows that the test characteristics curves of the four test forms overlap one another. This indicates that in general the four listening test forms of the 2017 TOEP are equivalent with one another. However, more detailed examination and more attention to the ability scales of the test takers reveal that the curves overlap one another for the test takers of the ability scale of around

0-1, while for the test takers of the ability scale <0, the curves is seen a little loose, so are the curves for the test takers of ability scale >1. From this, it can be understood that for the test takers of the ability scales of <0 and >1, the test forms are rather inequivalent in terms of difficulty level, though the difference is not dominant because visually the distance of the curves for the ability scales is still very close.

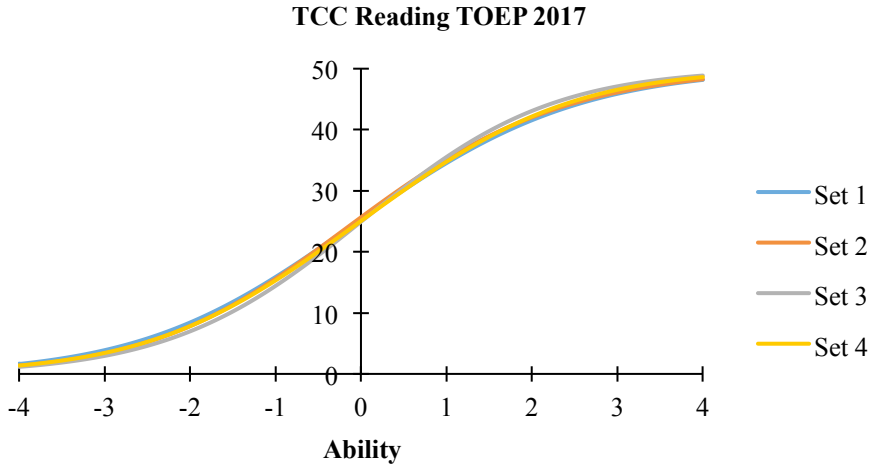

Figure 4. The Analysis of the Equivalence of the Four Reading Test Forms of the 2017 TOEP

The same phenomenon is seen in Figure 4, in which the characteristics curves of the four reading test forms also overlap one another, though the curves are rather loose for test takers of the ability scale <0 and >1. This indicates that in general the four reading test forms of the 2017 TOEP are equivalent with one another. In this way, the evidence of the equivalence of the four test forms of the 2017 TOEP simultaneously proves that the difficulty indices of the four listening and reading test forms of the 2017 TOEP are also relatively equivalent so that no test takers are disadvantaged. Such an evidence also ensures that the scores obtained by the test takers are empirically comparable.

Discussion

When a test is developed in several parallel forms, it is very important to ensure the equivalence of the test forms (Baghaei, 2010; Kartowagiran et al, 2018; Retnawati, Kartowagiran, Arlinwibowo, & Sulistyaningsih, 2017; van der Linden, 2013; von Davier & Wilson, 2007). This is to ensure the test forms used are fair for all test takers (Baghaei, 2010; Kartowagiran et al., 2018; Meyer & Zhu, 2013; Retnawati et al., 2017). The results of analysis indicate that the listening and reading test forms of both the 2016 TOEP and 2017 TOEP are equivalent. The proven equivalence of the test forms becomes one of the indicators that the test forms have a good quality (Aşiret & Sūnbūl, 2014). There are several reasons why the equivalence of the different forms is important. First, the equivalent test forms are fairer for test takers. The equivalence indicates that the aspect being measured is the same and the test item parameter is also the same so that no test takers are disadvantaged. In addition, with the equivalent test forms the scores obtained by different test takers can be compared. Second, the use of parallel test forms to measure the same ability can minimize cheating during the test (Meyer & Zhu, 2013) because when different but parallel test forms are used, the test items can be jumbled. In this way, although the test takers are doing the test at the same time, the test items might be different. Third, it is for security reason, i.e. to protect the confidentiality of the test items. When the test has equivalent forms, the answer key can be kept confidential because the test items can be different across test times although the same aspects are measured.

Apart from indicating that the test forms have a good quality, the evidence of the test form equivalence also indicates that the test forms have been developed based on the same grid. A grid has a crucial role in the construction of test items (AlFallay, 2018; Cohen & Wollack, 2003; Fives & DiDonato-Barnes, 2013), especially the development of a large number of items (item bank). The grid will help reveal the scope of materials to be tested. The test grid will help test developers to consistently measure the same ability, although the items are different. In this way, the test grid can help test developers to construct the desired test items which can consistently measure the same ability though the test items are distributed in different test forms.

The evidence of the equivalence of the listening and reading test forms of the 2016 TOEP and the 2017 TOEP indicates that the construct used to construct the test is the same. A construct is an attribute, competency, ability,

or skill occurring in the human brain and defined by the established theorie (Brown, 2000) and contain an inductive summary (Cronbach & Meehl, 1955) of the measured aspects. This indicates that a construct is a theoretical framework of the test to develop by test developers. By using the same construct, the same framework is automatically used to construct test items for different forms. In addition, the test construct has been proven to have a high level of validity, in which the purpose of investigating the construct validity is to find out whether the theoretical construct used to develop the test is consistent with the empirical construct. In this way, the evidence of the construct validity ensures that the theoretical construct used to develop the test is reliable and can be used consistently in the process of developing test forms.

Besides the grid and test construct factors, another factor which contributes to the development of parallel TOEP forms is the developers with their experience in developing test items. Experience in developing test items is badly needed to produce quality test items so that the test results can give advantages to different parties, such as, instructional designers, teachers, students, and test administrators (Liu, 2017). Experienced test developers can easily understand the test grid so that the test items they have constructed have a high degree of relevance for the ability to be measured. In addition, with the experience they have, test developers are theoretically more able to consider various aspects related to item characteristics, such as difficulty levels, differentiating power, and the function of distractors. Items which are either too easy or too difficult are to be revised for future use by considering the blueprint of the item cores.

CONCLUSIONS

Based on the findings and discussion, it can be concluded that the six listening and reading test forms of the 2016 TOEP and the four listening and reading test forms of the 2017 TOEP are proven to be equivalent. This finding has provided evidence that the test forms have a good quality and are fair for all test takers. Some points which support the quality of the test forms are as follows: (1) the test forms have been developed by using the same grid; (2) the test forms have been developed based on the same construct; and (3) the test items have been constructed by a team of experienced developers. The followings are some recommendations related to the finding of the research study. *Firstly*, to develop a quality test, especially in the case of a test

consisting of different parallel forms, the developers need not only to analyze item characteristics through the classical or modern approach, but also to analyze the equivalence of the parallel test forms. *Secondly*, an item bank needs to be developed based on the test grid and construct already tested, especially for assessment in the fields of education and psychology. *Thirdly*, proving the equivalence of test forms is not limited to the use of the Racsh Model. Test analyses using the model 2 of logistic parameter and model 3 of logistic parameter are likely to provide richer information concerning the quality of the test forms.

REFERENCES

Aşiret, S., & Sūnbūl, S. O. (2014). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice, 16*(2), 647-668.

AlFallay, I. S. (2018). Test specifications and blueprints: Reality and expectations. *International Journal of Instruction*, *11*(1), 195-210.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling*, *52*(3), 313-322.

Bartram, D. (2008). The advantages and disadvantages of on-line testing. In S. Cartwright & C. L. Cooper (Eds.)*, The Oxford handbook of personnel psychology* (pp. 234-260). Oxford: Oxford University Press.

Brookhart, S. M. & Nitko, A. J. (2015). *Educational assessment of students (7th ed.)*. Boston, MA: Pearson Education.

Brown, H. D. (2004). *Language assessment principles and classroom practices.* New York, NY: Pearson Education.

Brown, J. D. (2000). What is construct validity? *JALT Testing & Evaluation SIG Newsletter, 4*(2), 8-12.

Cohen, A. S., & Wollack, J. A. (2003). Helpful tips for creating reliable and valid classroom tests: Getting started—The test blueprint. *The Learning Link, 3*(4), 1-2.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. London: Psychology Press.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research & Evaluation, 18*(3), 1-7.

Foster, P. (2009). *Teacher's guide for English for today: Grade-VIII*. Dhaka, Bangladesh: National Curriculum and Textbook Board.

Gitsaki, C., & Robby M. A. (2018). Benefit of language assessment. In J. I. Liontas (Eds.), *The TESOL encyclopedia of English language teaching I* (pp. 1-6). Hoboken, NJ: John Wiley & Sons.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Kartowagiran, B., Munadi, S., Retnawati, H., & Apino, E. (2018). The equating of battery test packages of Mathematics national examination 2013-2016. *SHS Web of Conferences, 42*(22), 1-6.

Kolen, M. J., & Brenann, R. L. (2004). *Test equating: Methods and practice*. New York, NY: Springer.

Linn, R. L., & Gronlund, N. E. (2005). *Measurement and assessment in teaching*. Singapore: Pearson Education.

Liu, Y. (2017). Assessment in English language teaching and learning. *Proceedings of the 3rd International Conference on Education and Social Development (ICESD),* 127-131.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equation. *Research & Practice in Assessment, 8*, 26-39.

Norris, J. M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum, 38*(1), 41-45.

Pang, X., Madera, E., Radwan, N., & Zhang, S. (2010). A Comparison of Four Test Equating Methods. Report prepared for the Education Quality and Accountability Office (EQAO). Retrieved July 4, 2012.

Rahman, M. F., Babu, R., & Ashrafuzzaman, M. (2011). Assessment and feedback practices in the English language classroom. *Journal of NELTA, 16*(1-2), 97-106.

Retnawati, H. (2014). *Teori respons butir dan penerapannya [Item response theory and its application]*. Yogyakarta, Indonesia: Parama Publishing.

Retnawati, H. (2016). *Perbandingan metode penyetaraan skor tes menggunakan butir bersama dan tanpa butir bersama [Comparisons of test score equating methods using shared items and without shared items]*. *Jurnal Kependidikan, 46*(2), 164-178.

Retnawati, H., Kartowagiran, B., Arlinwibowo, J. & Sulistyaningsih, E. (2017). Why are the Mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction*, *10*(3), 257-276.

Ryan, J., & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on classical test theory and item response theory*. Technical Issues in Large-Scale Assessment (TILSA), State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO). Retrieved from https://files.eric.ed.gov/fulltext/ED544690.pdf

Sawe, B. E. (2019). *What is the most spoken language in the world?* Retrieved from https://www.worldatlas.com/articles/most-popular-languages-in-the-world.html

Smith, R. M., & Kramer, G. A. (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement, 52*(4), 835-846.

Sutrisno, H. (2016). An analysis of the Mathematics school examination test quality. *Jurnal Riset Pendidikan Matematika, 3*(2), 162-177.

van der Linden, W. J. (2013). Some conceptual issues in observed-scored equating. *Journal of Educational Measurement, 50*(3), 249-285.

von Davier, A. A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement, 67*(6), 940-957.

West, C. (2010). Borderless via technology. *International Educator*, *19*(2), 24

Yim, M. K., & Huh, S. (2006). Test equating of the medical licensing examination in 2003 and 2004 based on the item response theory. *Journal of Educational Evaluation for Health Professions, 3*(2), 1-5.

Yu, C. H., & Popp, S. E. O. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment, Research & Evaluation*, *10*(4), 1-19.