ANIMA Indonesian Psychological Journal
2020, Vol. 35, No. 2, 133-156
https://doi.org/10.24123/aipj.v35i1.2905
P-ISSN: 0215-0158 E-ISSN: 2620-5963

Testing of the Indonesian Version of the Instrument Teachers' Sense of Efficacy Scale Using Rasch Modelling

Herdiyan Maulana, Anna Armeini Rangkuti, and Lussy Dwiutami Wahyuni
Fakultas Pendidikan Psikologi
Universitas Negeri Jakarta

Bambang Sumintono
Institute of Educational Leadership
Universiti Malaya

Teachers who have self-efficacy regarding their abilities play an important part in the educational process. The instrument "Teachers' Sense of Efficacy Scale (TSES)", which consists of 24 items, measures the self-efficacy of a teacher in conducting their role in the classroom. This research was aimed at testing the psychometric properties of the Indonesian language version of the TSES, using Rasch testing. 245 State Primary School teachers in the *Daerah Khusus Ibukota* (*DKI* - Special Capital District) Jakarta, Indonesia, took part in the study. Analysis results indicated that the reliability index and the separation of item and person scales of the TSES fulfil the criteria well. In the study of the structure of the organization of the instrument, tests of the unidimensional quality and accuracy of the items/models showed they fulfil the conditions for measurement by Rasch modelling. In the rating scale, data analysis indicated the necessity of simplification of the range of response choices, and also of the variation in the difficulty of the items, which is not very uniform, so that a large proportion of respondents were not easily measurable, via this instrument. The results of the research indicated two items which showed bias, when dealt with by different groups of respondents. Based upon the information resulting from this testing, this script discusses recommendations for the modification of the Indonesian language version of the TSES instrument, with the aim of producing an instrument having good psychometric qualities, to measure the self-efficacy of teachers.

Keywords: instrument quality, Rasch modelling, Teachers' Sense of Efficacy Scale

Received 5 August 2017; Accepted 3 January 2018; Published 25 April 2020.

Teachers' efficacy (TE) is the self-conviction of a teacher regarding their ability to execute their duties as an educator (Tschannen-Moran & Hoy, 2001). In the past decade, the TE concept has continued to attract the attention of researchers in a number of countries, focused upon the testing of the psychometric quality of the TE instrument (Kleinsasser, 2014; Mahdinezad & Mansouri, 2016). The root of the TE concept is the social cognition theory (Bandura, 1994), particularly concerning triadic reciprocal causation, which explains that human behavior is the product of the interaction between internal and external factors. Internal factors cover the personal domain, such as the attitudes and behavior of the individual, whilst external factors refer to a number of externals, such as the environment and the behavior of others. This article will clarify the analysis of the instrument for the measurement of a TE, adopted from the Teachers' Sense of Efficacy Scale (TSES; Tschannen-Moran, 2004), using the Rasch modelling scale approach. The portion following that explains the development of the TE concept, followed again by an explanation of the methodology used, discussion of the results of the research, and finally the conclusions of the research.

The TE concept was developed by Maddux (1995) and developed further by Soddak and Poddel (1996), stressing the association between the TE concept and the theory of social cognition in the context of the internal locus of control. This internal locus of control explains that, basically, human kind has the cognitive capacity to engage in reflection and self-deve-

Correspondence concerning this article should be addressed to Herdiyan Maulana, Fakultas Pendidikan Psikologi, Universitas Negeri Jakarta, Jl. R. Mangun Muka Raya, DKI Jakarta 13220. Email: herdiyanmaulana@gmail.com

lopment, within the framework of aligning themselves with their constantly changing social environment, based upon awareness that the individual has full control of all of these aforesaid things. Thus, the efficacy of a teacher is the process of the formation of attitudes and behavior, influenced by these said internal and external factors, in accord with their efforts to align themselves with the environment, based upon awareness of the duties of a teacher.

TE is one of the important elements in determining the degree of achievement of the goal of the lesson (Klassen et al., 2011; Yeo et al., 2008). A number of the most recent pieces of research have indicated that TE has impacts upon the performance of a teacher in class. Several of these are the raising of the effectiveness of the management of classes having students of culturally diverse backgrounds (Tucker et al., 2005), the maintenance of a commitment to professionalism, and the level of teacher retention in a school (Darling-Hammond, 2003; Ware & Kitsantas, 2007), as well as assisting the teacher to give good lesson instruction. (Dixon et al., 2014). A number of these pieces of research underline several benefits of TE for the teacher in general, i.e., assisting the quality of instruction in the class.

Furthermore, TE has several functions in individual and organizational upgrading. In the personal context, TE has the potential to optimize the ability of a teacher to execute their duties as an educator, amongst which are making possible the opening of a variety of concepts, and, indeed, new breakthroughs in teaching methods (Tschannen-Moran & Hoy, 2001). Furthermore, TE also has a part in assisting a teacher in taking the steps selected to perform instructional activities, as well as the perseverance of the teacher in overcoming impediments and failures in teaching, and to elicit resilience in the teacher to anticipate instructional challenges in class (Tschannen-Moran & Gareis, 2004; Wilcox-Herzog & Ward, 2004).

Meanwhile, in the overall organizational grading of the school, teachers with high TE ratings have a correlation with a healthy, ordered and positive school organizational climate, as well as increasing the possibilities of accurate decision making by teachers in the classroom in which they teach, and also the collective efficacy of all elements in the school (Tschannen-Moran, 2003).

The measurement of TE was initiated by Bandura (1977) with the introduction of four efficacy indicators for teachers, i.e.,: (1) achievement of the goals of the learning process (performance accomplishment); (2) direct experience (vicarious experience); (3) verbal encouragement (verbal persuasion); and (4) emotional stimulation (emotional arousal). These four aforesaid dimensions became the basis for the composition of the TE instrument by Bandura (1977). Tschannen-Moran and Hoy (2001) consistently tested this instrument, which later gave rise to a proposed instrument, called the Ohio State Teacher Efficacy Scale (OSTES). Subsequently, the concept of the OSTES was ultimately perfected, with the publication of research concerning the TSES, compiled by Tschannen-Moran (2004). This scale consists of a nine-point evaluation continuum (rating scale) the wording of which always begins with, "*How good is your ability to...?*", at each item. The categories of the responses consist of: "1 (*unable to*)", "3 (*little ability*)", "5 (*adequate ability*)", "7 (*great ability*)", and "9 (*very great ability*)". This scale is also provided in a short format, and a long/unabridged format, based upon the number of items. In the short format, the number of items is 12, whilst in the unabridged format, it is 24 items, including the 12 items in the short format. Each item comprises the results of the reduction of three dimensions, considered representative of teaching duties in general, i.e., (1) student engagement (the duties of the teacher in increasing the involvement of students in the activities of the school); (2) instructional strategies (the duties of the teacher in applying effective teaching strategies); and (3) classroom management (the duties of the teacher in the management of the class). The scores in each statement in the TSES are divided equally and have no ranking (*hierarchy)*.

The testing of the validity and reliability of the TSES continued to develop, and indicated that testing of the instrument, with the variation of methods and characteristics of the varied forms of samples, was still required. (Kleinsasser, 2014). The TSES scale has been used previously in research related to the self-confidence of instructors/teachers in a number of countries (Klassen & Chiu, 2010). A number of overseas studies have indicated that, although this scale has three different dimensions (student engagement, instructional strategies, and classroom management), the total of the scores from all of the statements are used to determine the TE levels. This was well confirmed to be this way, when tested in Western countries, such as in the US (Duffin et al., 2012), as well as in Asian countries, such as in Singapore (Nie et al., 2012). Although this is the case, other studies, such as that conducted by Scherer et al. (2016) concerning the testing of the quality of the TSES, using the Exploratory Structural Equation Modelling (ESEM), indicated that there is an *overlap* between several TE indicators. Using more

complex testing methods, and being concerned with the possibility of differences in the groups of samples, it was recommended that further testing be conducted to verify the psychometric qualities of this measuring implement. The testing of the psychometric qualities of the Indonesian version of the TSES, using a more complex psychometric analysis, has never previously been undertaken in Indonesia. Thus, it was thought necessary to fill this gap in knowledge, by presenting an analysis with an Item Response Theory approach, using Rasch modelling as a psychometric basis, for the use of the TSES in future years.

Rasch modelling can overcome the theoretical weaknesses of classical tests, related to pieces of research in the humanities, health, educational and social sciences fields, related to ordinal data arising from the perception/opinion/attitude of the respondents (Andrich, 1988; Bond & Fox, 2015). Rasch modelling can return data in accord with natural conditions, i.e., by using a probability approach. The odd probability values of raw ordinal data is sought, then calculated, based upon the Rasch logarithm, with the aim of producing measurements at equal distances (equal interval scale; Boone et al., 2014; Engelhard, 2013). At this stage, Rasch modelling accommodates a probability approach by looking at attributes as objects of measurement. This accommodation causes Rasch modelling not to have *deterministic* characteristics, so that it is able to identify the object of measurement more accurately (Sumintono & Widhiarso, 2013; 2015). Rasch analysis of TSES which have been conducted on the English language (Chang & Engelhard Jr, 2016) and the Malaysian language versions (bin Khairani & Razak, 2012). of the TSES, indicate that there is proof that the structure of the instrument and the reliability of the measurement implement, i.e., the *infit mean square* score, is between 0.75 to 1.25, and the Cronbach's Alpha value is 0.94., although several statements on the aforesaid scales were perceived as being inconsistent by several different respondents (the Differential Item Functioning [DIF] was low).

The testing of the TSES using Rasch modelling was to make possible the availability of more accurate information, and this testing process had not previously been conducted upon the Indonesian language version of the TSES. Thus, this conducting of Rasch modelling might also overcome problems of differences in measurement metrics between items. The calibration brought about by the Rasch modelling places the items and the subjects within the same measurements (Engelhard, 2013; Boone et al., 2014). The scores produced are no longer raw scores, with unknown measurement errors, but rather measurement scores with logit units (logarithm odd unit) not known certainly to have measurement *errors* in every item (Sumintono & Widhiarso, 2013; 2015). Testing using the Rasch methods, it was hoped, would give strong empirical evidence regarding the use of this scale in a wider context.

This research was aimed at conducting an evaluation of the quality of the TSES instrument, by using analysis based upon Rasch modelling, i.e., the Rasch rating scale analysis, to obtain psychometric information concerning the instrument, from scoring data on each item of the TSES. For that reason, with the Rasch method approach, the hypotheses were as follows.

Hypothesis 1: The TSES scale has unidimensional characteristics.

Hypothesis 2: Scale rating choices in the TSES function effectively.

Hypothesis 3: Items in the TSES scale can effectively illustrate the level of respondent variations.

Method

As a first step, the TSES instrument in English was translated into Indonesian by the researchers, using the back-translation procedure (Brislin, 1970). This method was undertaken by translating the English language version of the instrument into Indonesian, then translating that back into English, to look at the consistency of the translation results. This involved the *Unit Pelayanan Bahasa* (Language Services Unit), of the *Universitas Negeri Jakarta* (*UNJ*). In the back-translation process, language experts evaluated the re-translation results as being relatively the same as the original version. The blueprint of the Indonesian language version of the Indonesian language version of the TSES is presented in Table 1.

The codes for the dimensions used in the data analysis process are: "E = engagement"; "I = instructional"; and "C = classroom management". For example, for Item 1, the code is E21, meaning, E is for engagement, 2 is for the second indicator (able to overcome impediments), and 1 is for the first item in the sequence of E2, and so on. The rating scale used comprised the scores of 1 to 9, as used in the original English language questionnaire. The instrument was also furnished with questions for demographic variables, i.e., the gender, highest level of education, and public ser-

Table 1
Blueprint of the Bahasa Indonesia Version of the TSES

Factors	Indicators	Items	*Total Items*
(a) Efficacy in raising the involvement or motivation of students in learning activities *(student engagement)*	(1) Able to confront challenges related to endeavours to raise the level of involvement of students in the learning process, in class. (2) Able to overcome impediments related to endeavours to raise the level of involvement of students in the learning process, in class.	2, 6, 9, 12, 22 1, 4, 14	8
(b) Efficacy in conveying learning material or in knowledge-transfer duties *(instructional strategies)*	(1) Able to confront challenges related to conveying learning materials to children (2) Able to overcome challenges related to the conveyance of learning material to children	10, 11, 18, 20, 23, 24 7, 17	8
(c) Efficacy in class management *(classroom management)*	(1) Able to confront challenges related to class-management duties (2) Able to confront challenges related to class-management duties	5, 8, 13, 16 3, 15, 19, 21	8

vice status of the respondents. The Indonesian version of the instrument was then distributed to the respondents. The criterion for the respondents in this research was; had to be professional primary school teachers, living and teaching in the *Daerah Khusus Ibukota* (*DKI* - Special Capital District) Jakarta. The questionnaires returned by the respondents numbered 255, and all data was complete (no data was missing). The process of data selection (data filtering) was conducted, and there were 10 respondents showing the characteristics of outliers, because they had completed all answers with a score of 9, so their returns were later removed from the data. Therefore, the total number of respondents whose returns were analyzed was 245. Demographic information from the respondents is to be found in Table 2.

Table 2
Demographic Information from Respondents (N = 245)

Demographics	*n*	Percentage (%)
Gender		
Male	71	71%
Female	174	29%
Highest education level		
Diploma 3 (D3)	15	6%
Diploma 4 (D4)	8	3%
Bachelor (S1)	132	54%
Master (S2)	82	34%
Public service status of teachers		
Aspirant Public Servant (CPNS)	10	4%
Public Servant (PNS)	162	66%
Honorary*	73	30%

Note. *Employment status. Teacher paid by the government (known as "*pegawai negeri*") or by the individual school/organisation (honorary/ *honorer*).

The steps of data analysis were begun by entering the data from the questionnaires using Microsoft Excel (MS Excel) software, to be stored as data files in the .csv (Comma Separated Value) format, compatible with Rasch analysis. After the data was opened using the Winstep application, the psychometric analysis of the TSES was able to be performed. As explained above, Rasch analysis produces a conversion of the raw score from the TSES items, to logit scores. The logit scores obtained consist of those of the items and the those of the person (respondent). These scores were used in the subsequent analysis, in the Rasch modelling.

The first step in the analysis was to determine if the TSES scale had unidimensional characteristics. To determine the unidimensionality of the TSES, the writers looked at the results of the Winstep process, i.e., the Standardized Residual Variance, with a Principal Component Analysis (PCA) approach. The output results would indicate any variance values (raw variance), i.e., the minimal percentage of variance to be able to fulfil the standards of unidimensionality is 20%. The next step was analysis of the rating scales. This analysis was aimed at testing the validity of the choices for the scale ratings of the items, with the goal of confirming whether the choices of answers used were effective, and did not confuse the respondents. The observed average index values and those of the Andrich threshold were used as references to determine the effectiveness of these ratings. These values were seen as ideal, if the scores of the two indices underwent consecutive increases from their low negative values, upwards, towards positivity. This indicated that the choices of answer, for the ratings which were, given were effective, i.e., from low (negative) logit scores

going towards higher (positive) ones. Disorganization of the scores on the two indices was a sign of the presence of inefficient answer choice ratings, because not many were selected by the respondents.

The next analysis which was conducted was to evaluate the suitability of the items (item fit). The item fit analysis provided information concerning the level of difficulty of the items, in the Rasch analysis. This was aimed at seeing just how far an item had the tendency to attract positive (favorable) answers from the respondents. Rasch analysis via the Winstep application was able to produce a Wright map, which gave an illustration of the logit averages of the items, compared to the logit averages of the respondents. This analysis could also be seen in the Pearson measure score indicators of every item in the Winstep output, which examined in depth the item measures. If the average values were more than logit 0.00, this indicated that a correspondently increasing number of the items were answered positively by the respondents.

In every part of the item measures, a Cronbach's Alpha score could also be seen, which was able to be used as a reference for its reliability. Reliability is a measurement which illustrates the stability of an item, when used by different respondents at different times (Sumintono & Widhiarso, 2013). Reliability, in Rasch modelling, comprises two aspects: (1) the reliability of the item; and (2) the reliability of the person (respondent). The reliability scores hoped for were > .08 (Sumintono & Widhiarso, 2013). Quality testing of the items was conducted by referring to the Item Fit Order table. In this table, the information shows the ranking of the items, starting from the least appropriate to the data (item misfit). According to Boone et al. (2015), this is determined by three indices: (1) Outfit Mean-Square (MNSQ, in the range of 0.5 to1.5); (2) Outfit z-standardized (ZSTD, in the range of - 2 to + 2); and (3) point measure correlation (.4 to .85). If these values are not achieved, the said items may be categorized as not a fit (misfit).

Finally, an analysis of bias in the items was conducted by referring to the function of DIF. DIF in general functions to determine the difficulty of items. However, DIF also functions to consider whether items have a bias of not, in the category of certain respondents (Sumintono & Widhiarso, 2013). Items with a probability value below 5% (.05) would therefore be categorized as biased.

Table 3
Reliability of Persons and Items of the TSES Instrument

	Logit averages (SD)	*Separation*	*Reliability*	*Cronbach's Alpha*
Person	2.22 (1.46)	4.03	0.94	0.95
Item	0.00 (0.27)	2.72	0.88	

Results

Following is an explanation of the psychometric properties resulting from the processing, using Rasch modelling, of the TSES data obtained from 245 participants in this research. In general, the average logit values of the respondents indicated a tendency to be high (+ 2.22 logit). This showed that all of the respondents had a positive TE score, (i.e., higher than 0.00 logit, which is the logit standard for the level of item difficulty). This was made increasingly clear, with a deviation standard value of 1.46 logit, which increasingly indicated a logit range from a large proportion of the respondents (68%) being in the positive range (more than 0.00 logit). The average logit value of the items was 0.00 logit, which indicated a standard level of difficulty, however the standard deviation was .27, which showed a deficiency of uniformity in the level of difficulty of the TSES implement (see Table 3).

Evidence of Instrument Reliability

In the context of Rasch modelling, the indices used to measure the stability are the reliability and separation of the logit of items and persons, as is indicated in Table 3. Both the reliabilities of the respondents (person) and that of the items showed high values (.94 and .88, successively), and it was the same with the Cronbach's Alpha index values. The consistency of the analysis results using these two methods may be interpreted to mean the items of the TSES had consistency, when filled in by the research respondents.

The estimate of the parameters of the persons had a value of 4.03, which meant that the items might group the respondents, based upon their levels of efficacy, whilst an estimation of the parameters of the separations of the items had a value of 2.72 (rounded up to 3.00), meaning there were three categories of groups of respondents, based upon the ranking of their efficacy scores. With the separation value of the respondents being 4, the instrument used was able to map or group the respondents. This indicated that the respondents participating in the research were not homo-

genous, but rather heterogeneous, and showed they were a representative example of the population. Reliability information indicated that, although the aspects of the items and the persons were good, none the less those of the persons showed an index which was a little better.

Evidence of the Structure of the Compilation of the Instrument

The next step of the Rasch modelling concerns arguments in the matters of unidimensionality, appropriateness of the rating scales used, the index of the accuracy of the item model, (item fit statistics), as well as the existence of biased items (DIF). The results of each of these validity aspects, above, follow.

Scale of Unidimensionality. The results of the Standardized Residual Variance testing of the TSES instrument indicated that the raw variance had a value of 49.4%, which may be categorized as being satisfactory (Fisher, 2007). This indicated that the minimum variance of 20% had been met. Besides this, regarding the unexplained variance, this was between 3.0% and 4.6%, wherein nothing exceeded 15%, indicating that the instrument did indeed measure one variable which was TE.

Difficulty of the Rating Scale. The results of the analysis of the rating scale are shown in Table 4. The utility of the rating scale indicated only eight ratings were used (scores 2 to 9), with not one respondent having selected the score of 1 (*unable*). This indicates empirically that the ratings scale given should be slightly wider. On the observed average index, it was seen that there were values which rose and fell (- 0.43, then + 0.61, and falling to - 0.37) on ratings 2, 3 and 4. This indicated that the respondents were not really sure regarding the rating choices provided, that is on the scores of 2, 3 and 4. This was the case also with the Andrich threshold indications, which moved to - 0.70, then fell to - 2.16, and later rose to - 1.45, on the scores of 3, 4, and 5. These two matters indicated that the respondents did not really understand a number of rating choices given, particularly between the ratings with a score of 2 to 5. It was a different matter with higher scores, i.e., those of 6, 7, 8 and 9. Both the indices of the observed average and of the Andrich threshold indicated a rise in their values, which showed that the rating scales, in the view of the respondents, were separate, and could easily be differentiated. The graph of the scales in Figure 1 supports this assumption, because the plot of the probability of choosing a category, in the low categories, can be seen to be mutually overlapping, without any separation which could clearly be seen. This indicated that the separa-

Table 4
Statistics of the Analysis of the Ratings Scales

Rating (score)	*Observed Average*	*Andrich Threshold*
2	- 0.43	None
3	0.61	- 0.70
4	- 0.37	- 2.16
5	0.22	- 1.45
6	0,38	- 0.68
7	1.48	- 0.45
8	2.45	1.57
9	3.86	3.86

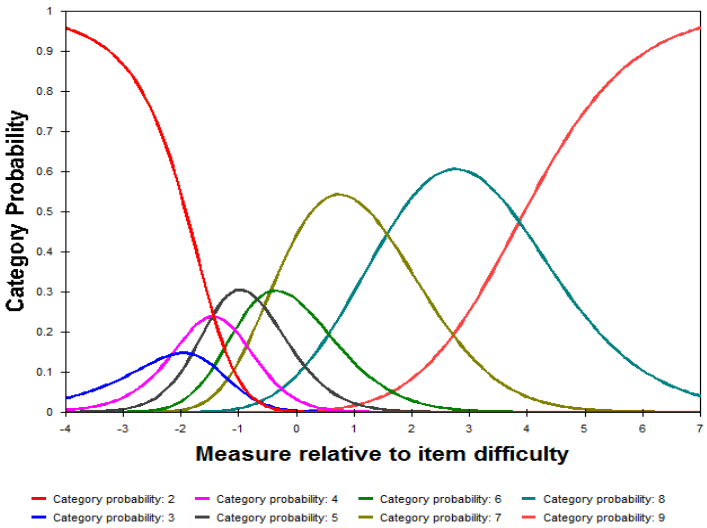


Figure 1. Analyses of the Rating Scales of TSES.

tion of the ratings was difficult to differentiate for the respondents, compared with the ratings with larger scores.

Item Fit. The results of the analyses of the suitability of the items is shown in Table 5, which is sequenced based upon the item number. The logit values show that there was no item which had too high a degree of difficulty (ease of being answered affirmatively by the respondents). The range of levels of difficulty of the items, or the items which could be answered in the negative, was between - 0.48 logit (Item 15 or C22) and + 0.43 logit (Item 22 or E14), wherein the number of respondents in this range was only 32 (13%) (see Figure 2). After that, the determination of the congruence between the item data and the model could be achieved by looking at the item fit statistics table (Table 5). As can be seen in Table 5, all items fulfilled the criteria of fit statistic for the Outfit MNSQ; whereas in the criteria for Outfit ZSTD there were five items which were outside the boundaries of accuracy (i.e., E22, E14, C14, I14 and E23). From the aspect of point measure correlation, the extant items met the criteria of being within the range of model item accuracy; with all values positive, this indicating that there was no item polarity in the TSES instrument. On the whole, there was no item which did not fulfil all three conditions, in the aspect of the validity of statistical suitability.

The Wright map (item/person map) in Figure 2 helps to explain the suitability between the items and the persons (respondents). The Wright map shows that the three dimensions of TE in the TSES instrument were within the range of being not too dissimilar, and at gradations easily understandable by the respondents. On this map, it may be seen that a number of respondents chose extreme values, i.e., on the right-hand edge of the map, located above the 'T' mark (meaning there was twice the standard deviation from the average), or above the logit value of + 5.00 logit. On the item aspect (left side of the map), it may be seen that C11 and C21 were in the same logit range. So it was with C22 and C24. On the dimension of the instructional engagement items, E12, E21 and E23 were also within the same logit range. This was also the case with the logit ranges of items I11, I12 and I14 in the instructional strategist, being on the same line.

Table 5
Statistics of Item Suitability (Item Fit Statistics)

No	Item	*Logit*	*Standard Error Measurement*	*Outfit MNSQ*	*Outfit ZSTD*	*Point Measure Correlation*
1	E21	+ 0.39	0.08	1.07	+ 0.70	0.71
2	E12	+ 0.36	0.08	0.82	- 1.86	0.71
3	C21	+ 0.18	0.09	1.14	+ 1.35	0.66
4	E22	+ 0.18	0.09	1.44	+ 3.95	0.61
5	C11	+ 0.14	0.09	1.04	+ 0.43	0.68
6	E22	- 0.18	0.09	0.98	- 0.12	0.65
7	I21	- 0.15	0.09	1.12	+ 1.15	0.64
8	C22	- 0.12	0.09	0.95	- 0.45	0.67
9	E13	- 0.30	0.09	1.07	+ 0.67	0.61
10	I11	- 0.20	0.09	0.99	- 0.06	0.63
11	I12	- 0.15	0.09	0.97	- 0.24	0.62
12	E14	- 0.05	0.09	0.88	- 1.24	0.66
13	C13	- 0.39	0.10	1.12	+ 1.15	0.66
14	E23	+ 0.31	0.08	0.72	- 3.07	0.75
15	C22	- 0.48	0.10	0.98	- 0.16	0.70
16	C14	+ 0.25	0.09	0.76	- 2.63	0.74
17	I22	+ 0.21	0.09	0.86	- 1.39	0.70
18	I13	+ 0.30	0.08	0.82	- 1.93	0.73
19	C23	- 0.43	0.10	0.97	- 0.26	0.66
20	I14	- 0.18	0.09	0.70	- 3.35	0.75
21	C24	- 0.19	0.09	0.99	- 0.06	0.69
22	E14	+ 0.43	0.08	1.28	+ 2.61	0.59
23	I15	+ 0.12	0.09	0.97	- 0.27	0.64
24	I16	- 0.05	0.09	1.02	+ 0.27	0.67

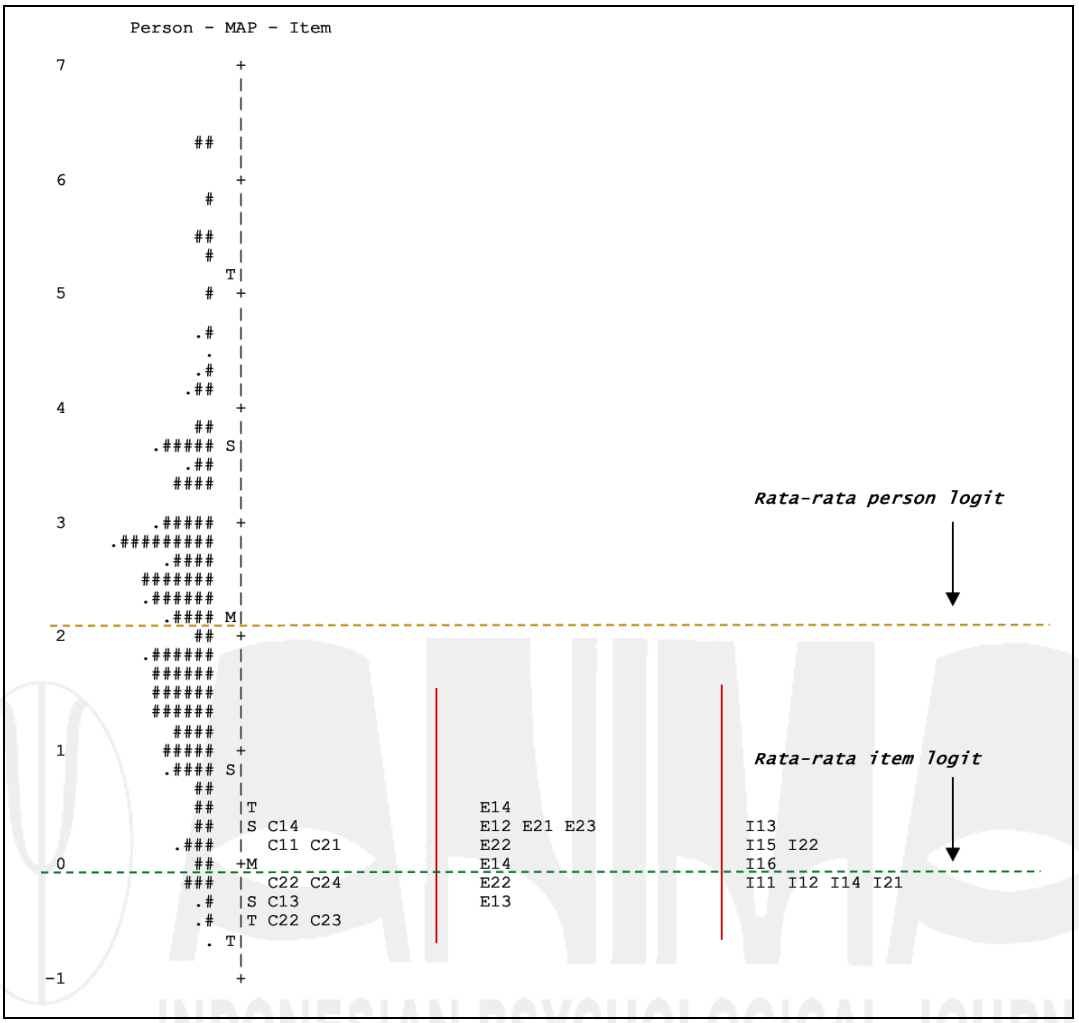


Figure 2. Wright Map for the TSES.

Note. C = Classroom Management; E = Instructional Engagement; I = Instructional Strategist.

DIF Analysis. The writers also conducted analysis to identify the items having differing functions, between the groups of the respondents, known as them being in the DIF category. There were three variants of variable demographic data which were used to determine the existence of biased items (gender, education and employment status). The results of DIF analysis indicated that there were two items which could be considered biased. Examined on the basis of gender, one was Item E21 (probability value of 0.0084), whilst on the basis of the demographic variable of employee status, Item I11 (probability value of 0.0427) also showed bias. There was no item which was biased on the basis of the response patterns, from the demographic variable of highest education level achieved.

Further information about the DIF analysis may be seen in Figure 3 and Figure 4. In Figure 3, what indicates a difference in the pattern of responses based upon the gender variable is that the black plot represents male teachers, whilst the red represents female teachers. When the plot nears the upper limits, this means the item has a high degree of difficulty, whilst when it nears the lower limits, this means the level of difficulty is low. For Item E21 (the item displaying DIF) in Figure 3, what is indicated is that because of a difference in the contrasting patterns of response (arrow), the item is considered easily answerable positively by male teachers, and at the same time is considered difficult to be answered positively by female teachers. This is in contrast to the item at the extreme

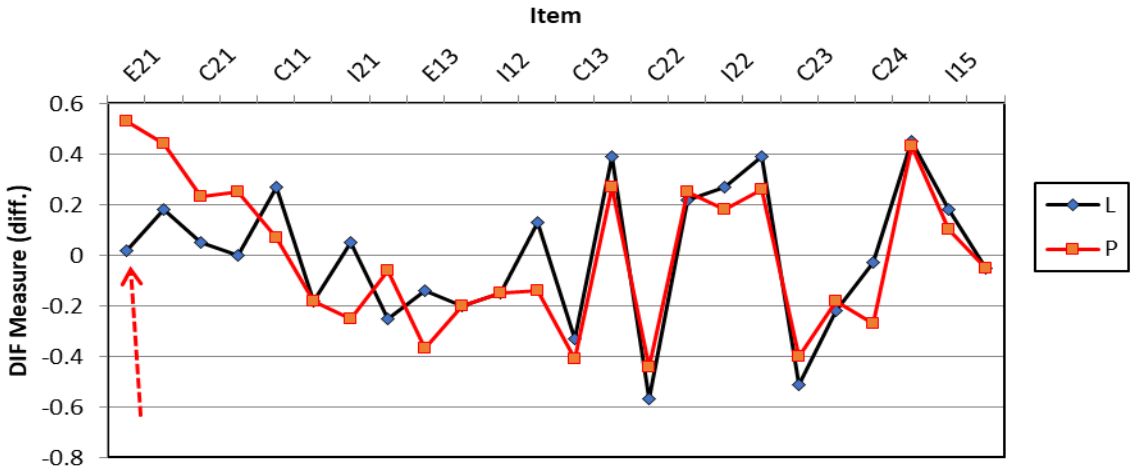


Figure 3. DIF analysis based upon the variable of gender.
Note. *L* = *Laki-Laki* (Male); *P* = *Perempuan* (Female).

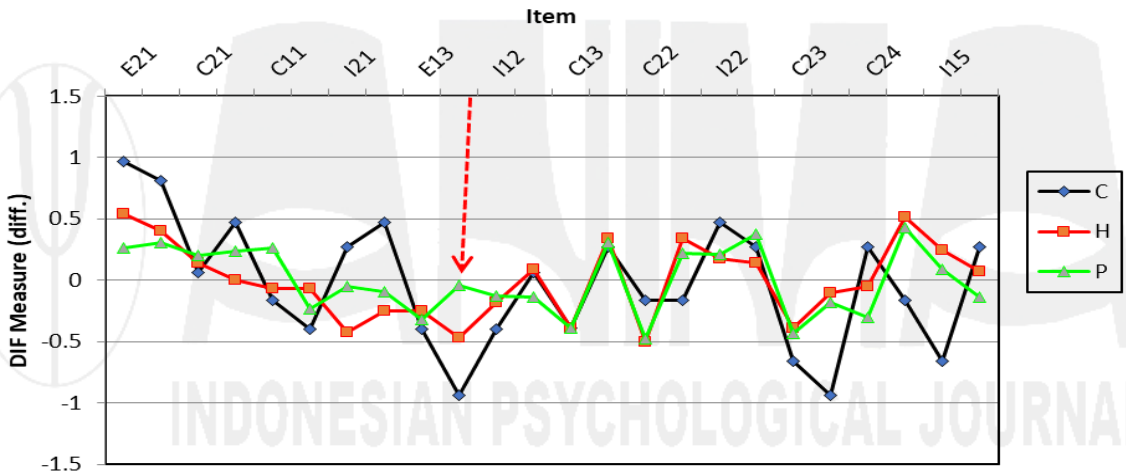


Figure 4. DIF analysis based upon the employment status variable.
Note. C = *CPNS*; H = Honorary; P = *PNS*

end, Item I15, where there is no difference in the response pattern between male and female teachers.

Figure 4 shows DIF analysis based upon employment status, consisting of three groups. Item I11, which has DIF (arrow), indicates differing response patterns for the three groups. The group of aspirant public servant teachers (*CPNS* [*Calon Pegawai Negeri Sipil*]; black plot) considered this item as one easily responded to positively, compared to the group of locally funded teachers (honorary) (paid by the school or organization) (red plot), however this item was more considered the most difficult to be responded to positively by the group of permanent public servant teachers (*PNS* [*Pegawai Negeri Sipil*]; green plot).

Discussion

Based upon the data analysis made by Rasch modelling, it may be seen that all of the respondents had a positive or high TE level (on average + 2.22 logit). The data analysis also indicated that the instrument used could map or group the respondents. The results of the mapping were that the respondents were not homogenous, but rather heterogenous. These results from the respondents involved in the research showed they were a reliable representation of the population. Besides this, the coefficient of reliability of the person indicated a very good level of respondent reliability.

Meanwhile, from the point of view of item quality, the results of the Rasch analysis indicated the level of

item difficulty tended towards little variability. This meant that a large proportion of the items were more easily answered positively by the respondents, so there was no answer variation. This made possible a state of social desirability. The 24 items used tended to be answered favorably by the respondents, so that the function of the items for the measurement of teacher efficacy needs to be reviewed; normally only 13% (32 respondents) could effectively be measured, based upon the possibility that the items were answered "*agree*" or "*do not agree*", by the respondents.

The first hypothesis of this research was supported, that the TSES had unidimensional characteristics. Based upon examination of the unidimensionality of the instrument, and the item/model accuracy (fit statistics), the TSES certainly measures only one variable, i.e., TE, and all items fulfil the conditions for statistical suitability, and all function uniformly. All items of the TSES were free of polarity (all point measure correlation values were positive).

The existence of a number of items in the same dimension, with a range of logit values which were not greatly different, indicated that the instrument can actually be further simplified, by the selection of only one of the items having the same function. Items with the same logit value indicate that these items measure the same concept (Bond & Fox, 2015; Boone et al., 2014). An example can be seen in Item 11 ("*How good are you at contriving good questions for the students*?" and Item 12 ("*How great are your endeavors to develop the students' creativity*?"). These two items measure the dimension of instructional strategy. Based upon the analysis results, the two questions not very efficiently, nor even qualitatively, measure the same concept, and certainly give the impression of over-lapping. Thus, what is needed is the revision of these items, i.e., it would be sufficient if only one of them was selected.

This indicates that an overhaul of the items should be performed, in all dimensions of TE in the TSES instrument. What could be done to improve this instrument is to alter or revise them, to become harder to answer in agreement, by the respondent (agreeable), especially when the logit values in one dimension are not too far apart. Another method which could be used is to alter several items to become unfavorable statements (negative statements), which, it may be assumed, would make the respondents think differently, and more seriously, as opposed to there being all favorable (positive) items. Despite this, from the viewpoint of randomization, the items work well.

The second hypothesis, related to the level of variation in the responses examined, based upon the scale rating choices, was not fully supported. In the Rasch analysis, the effectivity of the answer choices in the Likert scale (rating scale) is evaluated. In the TSES, although a rating scale of 1 to 9 is given, in fact the respondents chose from only the range of scores of 2 to 9, meaning that empirically the ratings, which for the respondents were effective, were only eight in number. At the same time, even with a rating scale of only eight, the range of choices by respondents, which could be categorized as low (rating scores 2 to 5), also did not work well. Thus, this indicated that the range of rating choices given was too large (a range of 1 to 9), and this situation had the potential to confuse the respondents. The solution is the simplification of the ratings, say from 1 to 4 or 1 to 5, with every rating choice being given a satisfactory explanation, such as is the case with the Likert rating scale (Bond & Fox, 2015). This situation can be caused by the different style of answering from each individual, influenced by individual feelings concerning the response forum provided in the instrument. There are individuals who feel their responses are covered by only two answer choices ("*Yes*" and "*No*"), however there are also those who feel it easier to answer when their choices have graduated options. Different preferences regarding the number of responses can produce structural differences which lower the quality of the information obtained from the completion of the instrument, in a survey (Widhiarso, 2016).

The final hypothesis concerning the level of variations from respondents, based upon the characteristics of the participants, was proven to be effective. This discovery was obtained from DIF analysis. The authors have made recommendations for improvement of the items which showed indications of bias, based upon *DIF* analysis, these being Item E21 ("*How good are your endeavors to respond to students who like to rebel*?") and Item I15 ("*How great are your endeavors to calm students who are disruptive and tumultuous*?"). These two items need revision, in order to prevent receiving differing answers from respondents, based upon their particular demographic categories (Boone et al., 2014). Qualitatively, these two items certainly had the potential to attract different responses. This may have been caused by perceptions regarding the differences in degrees of teacher authority, on the basis of their employment status, in considering the possibilities of answering in agreement with, or not in agreement with, the items. There should be deeper investigation into these assumptions. Although this is so, with only two items identified as having bias, this indicates that the quality of the items, on the whole,

showed not too many problems, when used with different categories of samples.

Aside from all of the discoveries of this research, the study had a number of notes on limitations which might be improved in the future. The first limitation was the sample size, which was relatively small, when compared with the sample population. Determination of appropriate measurements of the sample could assist in discovering variances in the items, based upon the characteristics of more uniform groups. According to Herrera and Gómez (2008), sample limitations in this experiment were particularly related to the interpretation of the DIF scores, when detecting if there was any bias in the items. Imbalances in the number of the samples of each sub-group would influence the degree of DIF accuracy, in detecting bias in the items (Herrera & Gómez, 2008) The next limitation was the process of adapting the instrument from the English language version, it being translated by the writers themselves, focusing upon grammar only, and not too concerned in the translation with examining the details of the context of the appropriateness of the items to the uniformity of the characteristics of the respondents. According to Cha et al. (2007), it is important to involve a number of people, from the group having the characteristics of the participants, in the translation process, in order to get an accurate translation result. On the basis of these limitations, it is important to make further analyses, using either a quantitative or qualitative method of approach, to examine the quality of the Indonesian version of the TSES, with the aim of verifying the results of the research conducted.

Conclusion

Based upon the analysis of the data, using Rasch modelling, the quality of the Indonesian version of the TSES instrument is, in many ways, quite good.

This instrument fulfils the basic principle of dimensionality, i.e., measuring only the variable of TE, and having only two biased items, as well as having 24 items without any polarity. On the other hand, there are several shortcomings which could be improved in the future, these being the low level of difficulty of items, so that it was not possible to determine the TE of respondents whose logit values are high (respondents who tend always to give answers which agree).

Several matters which need to be undertaken in the future, related to the improvement of the instrument, are the simplification of the rating scale (the range of choice intervals) needs to be made shorter, the revision of several items in one dimension to have a higher level of difficulty, as well the need for the modification of the formats of several items, into forms eliciting unfavorable answers. Further research could also improve the translation of the instrument into Indonesian, which could influence the comprehension of the participants and respondents to the TSES items. The translation of the TSES could involve language experts and individuals originating from the groups with the characteristics of the participants.

Pengujian Kualitas Instrumen *Teachers' Sense of Efficacy Scale* Versi Bahasa Indonesia Menggunakan Pemodelan Rasch

Herdiyan Maulana, Anna Armeini Rangkuti, and Lussy Dwiutami Wahyuni
Fakultas Pendidikan Psikologi
Universitas Negeri Jakarta

Bambang Sumintono
Institute of Educational Leadership
Universiti Malaya

Guru yang mempunyai efikasi diri akan kemampuannya merupakan bagian penting dalam proses pendidikan. Instrumen *Teachers' Sense of Efficacy Scale* (*TSES*), yang berisi 24 butir, mengukur efikasi diri seorang guru dalam menjalankan perannya di kelas. Penelitian ini bertujuan untuk menguji properti psikometrik *TSES* versi Bahasa Indonesia dengan menggunakan pemodelan Rasch. Sebanyak 245 orang guru SD (Sekolah Dasar) Negeri di DKI (Daerah Khusus Ibukota) Jakarta berpartisipasi dalam studi ini. Hasil analisis menunjukkan indeks reliabilitas serta separasi *item* dan *person* skala *TSES* memenuhi kriteria yang baik. Dalam kajian struktur penyusun instrumen, uji unidimensionalitas dan ketepatan butir-model memenuhi syarat dalam pengukuran pemodelan Rasch. Dalam hal skala peringkat (*rating scale*), analisis data menunjukkan perlunya penyederhanaan rentang pilihan jawaban; juga variasi tingkat kesulitan butir yang tidak terlalu beragam sehingga sebagian besar responden tidak dapat terukur dengan baik melalui instrumen ini. Hasil penelitian ini menunjukkan dua butir yang memiliki bias ketika di kerjakan oleh kelompok responden yang berbeda. Berdasarkan informasi hasil pengujian tersebut, tulisan ini membahas usulan modifikasi instrumen *TSES* versi Bahasa Indonesia dengan tujuan menghasilkan instrumen efikasi diri guru yang memiliki properti psikometrik yang baik.

Kata kunci: kualitas instrumen, pemodelan Rasch, *Teachers' Sense of Efficacy Scale*

Masuk 5 Agustus 2017; Terima 3 Januari 2018; Terbit 25 April 2020.

Teachers' efficacy (*TE*) adalah keyakinan diri guru akan kemampuannya dalam menjalankan tugas sebagai pendidik (Tschannen-Moran & Hoy, 2001). Dalam satu dekade terakhir, konsep *TE* terus menjadi perhatian para peneliti di berbagai negara dengan berfokus pada pengujian kualitas psikometri instrumen *TE* (Kleinsasser, 2014; Mehdinehzad & Mansouri, 2016). Akar konsep *TE* adalah teori kognisi sosial (Bandura, 1994), khususnya mengenai *triadic reciprocal causation* yang memberikan penjelasan bahwa perilaku manusia merupakan produk interaksi antara faktor internal dan eksternal. Faktor internal meliputi domain personal, seperti sikap dan perilaku individu. Sementara faktor eksternal mengacu pada berbagai hal di luar individu, seperti lingkungan dan perilaku orang lain. Artikel ini akan menjelaskan analisis intrumen untuk mengukur *TE* yang diadopsi dari *Teachers' Sense of Efficacy Scale* (*TSES*; Tschannen-Moran & Gareis, 2004), dengan menggunakan pendekatan pengukuran pemodelan Rasch. Bagian selanjutnya menjelaskan perkembangan konsep *TE*, dilanjutkan dengan penjelasan metodologi yang digunakan, pembahasan dan diskusi hasil penelitian, dan di akhir dengan kesimpulan riset.

Konsep *TE* dikembangkan oleh Maddux dan Lewis (1995) dan lebih lanjut oleh Soodak dan Poddel (1996) dengan menekankan keterkaitan antara konsep *TE* dengan teori kognisi sosial dalam konteks rentang kendali internal (*internal locus of control*). Rentang kendali internal ini menjelaskan bahwa pada dasarnya manusia memiliki kapasitas kognisi untuk melakukan refleksi dan pengembangan diri dalam rangka menyesuaikan diri terhadap lingkungan sosialnya yang terus berubah yang dilandasi atas kesadarannya bahwa individu memiliki kontrol penuh terhadap semua hal tersebut. Dengan demikian, efikasi guru

Korespondensi sehubungan dengan artikel ini ditujukan pada Herdiyan Maulana, Fakultas Pendidikan Psikologi, Universitas Negeri Jakarta, Jl. R. Mangun Muka Raya, DKI Jakarta 13220. Email: herdiyanmaulana@gmail.com

adalah proses pembentukan sikap dan perilaku yang dipengaruhi oleh faktor internal dan eksternal tersebut, seiring dengan upaya untuk menyesuaikan diri dengan lingkungan berbasis keyakinannya atas tugas-tugas sebagai guru.

TE merupakan salah satu elemen penting dalam memastikan tercapainya tujuan pembelajaran (Klassen et al., 2011; Yeo et al., 2008). Beberapa penelitian terkini menunjukkan bahwa *TE* memiliki dampak terhadap kinerja guru dalam kelas, beberapa diantaranya adalah peningkatan efektifitas pengelolaan kelas dengan latar belakang siswa yang memiliki budaya berbeda (Tucker et al., 2005), mempertahankan komitmen profesionalitas dan tingkat retensi guru di sekolah (Darling-Hammond, 2003; Ware & Kitsantas, 2007), serta membantu guru dalam memberikan instruksi pembelajaran dengan baik (Dixon et al., 2014). Beberapa penelitian tersebut menggarisbawahi beberapa manfaat *TE* terhadap kinerja guru secara umum, yaitu membantu meningkatkan kualitas pembelajaran di dalam kelas.

Lebih lanjut, *TE* memiliki beberapa fungsi dalam tataran individu dan organisasi. Dalam konteks personal, *TE* berpotensi untuk mengoptimalkan kemampuan guru dalam menjalankan tugasnya sebagai pendidik, diantaranya ialah memungkinkan terbukanya berbagai gagasan, bahkan terobosan baru dalam metode pengajaran (Tschannen-Moran & Hoy, 2001). Lebih lanjut, *TE* juga turut membantu guru dalam mengambil tindakan yang dipilih guru dalam melakukan kegiatan instruksional, serta kegigihan guru dalam mengatasi hambatan dan kegagalan dalam mengajar, serta memunculkan resiliensi/daya tahan guru dalam mengantisipasi tantangan pembelajaran di kelas (Tschannen-Moran & Gareis, 2004; Wilcox-Herzog & Ward, 2004).

Sementara itu pada tataran organisasi sekolah secara keseluruhan, guru dengan tingkat *TE* yang tinggi berkorelasi dengan iklim organisasi sekolah yang sehat, teratur dan positif, serta meningkatnya kemungkinan pengambilan keputusan yang tepat oleh guru di ruang kelas tempatnya mengajar, dan juga tingginya tingkat keyakinan bersama (*collective efficacy*) pada semua elemen di sekolah (Tschannen-Moran & Hoy, 2001).

Pengukuran *TE* pertama kali diinisiasi oleh Bandura (1977) dengan memperkenalkan empat indikator efikasi guru, yaitu: (1) pencapaian tujuan pembelajaran *(perfomance accomplishment)*; (2) pengalaman langsung *(vicarious experience)*; (3) dorongan verbal *(verbal persuasion);* dan (4) stimulasi emosi *(emotional arousal)*. Keempat dimensi tersebut menjadi landasan dalam penyusunan instrumen *TE* oleh Bandura (1977). Tschannen-Moran dan Hoy (2001) secara konsisten menguji instrumen ini, yang kemudian bermuara pada usulan model instrumen dengan nama *Ohio State Teacher Efficacy Scale* (*OSTES*). Lebih lanjut, konsep *OSTES* kemudian disempurnakan dalam kurun waktu terakhir dengan dipublikasikannya riset tentang *TSES* yang disusun oleh Tschannen-Moran dan Gareis (2004). Skala ini terdiri dari sembilan poin kontinum penilaian (peringkat/*rating scale*) yang selalu diawali dengan pertanyaan "*Seberapa besar kemampuan Anda...?*" pada setiap butirnya. Kategori penilaiannya terdiri dari: "1 (*tidak mampu*)", "3 (*kemampuan kecil*)", "5 (*kemampuan cukup*)", "7 (*kemampuan besar*)", dan "9 (*kemampuan sangat besar*)". Skala ini juga tersedia dalam format singkat dan format panjang/utuh, yaitu berdasar banyaknya butir. Skala dalam format singkat terdiri dari 12 butir. Sedangkan skala dalam format panjang terdiri dari 24 butir, yang termasuk 12 butir yang ada pada skala dalam format singkat. Setiap butir merupakan hasil penjabaran dari tiga dimensi yang dianggap merepresentasikan tugas-tugas pengajaran secara umum, yaitu: (1) *student engagement* (tugas guru dalam meningkatkan keterlibatan siswa terhadap aktivitas sekolah); (2) *instructional strategies* (tugas guru dalam menerapkan strategi pengajaran efektif); dan (3) *classroom management* (tugas guru dalam manajemen kelas). Skor pada setiap pernyataan di *TSES* dibagi rata dan tidak memiliki jenjang (*hierarchy*).

Pengujian validitas dan reliabilitas *TSES* terus berkembang dan menunjukan bahwa pengujian instrumen dengan variasi metode dan karakteristik sampel yang beragam masih dibutuhkan (Kleinsasser, 2014). Skala *TSES* telah digunakan sebelumnya pada penelitian terkait keyakinan diri pengajar/guru di berbagai negara (Klassen & Chiu, 2010). Beberapa studi di luar negeri menunjukkan, walau skala ini memiliki tiga dimensi yang berbeda (*student engagement*, *instructional strategies*, dan *classroom management)*, skor total dari keseluruhan pernyataan digunakan untuk melihat tingkat *TE*. Hal ini dikonfirmasi baik itu ketika diuji di negara barat, seperti Amerika Serikat (Duffin et al., 2012), maupun di negara Asia, seperti Singapura (Nie et al., 2012). Namun demikian, studi lainnya, seperti yang dilakukan oleh Scherer et al. (2016) tentang pengujian kualitas *TSES* dengan menggunakan *Exploratory Structural Equation Modeling* (*ESEM*) menunjukkan adanya *overlapping* pada beberapa indikator *TE*. Penggunaan metode tes yang lebih kompleks dengan memperhatikan kemungkinan perbedaan pada kelom-

pok sampel direkomendasikan untuk dilakukan lebih lanjut untuk memverifikasi kualitas psikometri alat ukur ini. Pengujian kualitas psikometri *TSES* versi Indonesia dengan menggunakan analisis psikometri yang lebih kompleks belum pernah dilakukan di Indonesia. Dengan demikian, perlu untuk mengisi celah pengetahuan ini dengan menyajikan analisis dengan pendekatan teori *Item Response Theor*y dengan model Rasch sebagai landasan psikometri bagi penggunaan *TSES* di masa yang akan datang.

Pemodelan Rasch dapat mengatasi kelemahan teori tes klasik dalam kaitannya dengan penelitian-penelitian humaniora, kesehatan, pendidikan, dan ilmu sosial lain yang berhubungan dengan data ordinal yang berasal dari persepsi/opini/sikap responden (Andrich, 1988; Bond & Fox, 2015). Pemodelan Rasch dapat mengembalikan data sesuai dengan kondisi alamiahnya yaitu dengan menggunakan pendekatan probabilitas. Data mentah yang berjenis ordinal dicari nilai *odd probability*-nya, kemudian dihitung berdasarkan logaritma Rasch dengan tujuan menghasilan pengukuran dengan jarak yang sama (*equal interval scale*; Boone et al., 2014; Engelhard, 2013). Pada tahap ini, pemodelan Rasch mengakomodasi pendekatan probabilitas dalam memandang atribut sebagai sebuah obyek ukur. Pengakomodasian ini menyebabkan pemodelan Rasch tidak bersifat *deterministic,* sehingga mampu mengidentifikasi objek ukur secara lebih cermat (Sumintono & Widhiarso, 2013; 2015). Analisis Rasch terhadap *TSES* sudah dilakukan pada *TSES* versi Bahasa Inggris (Chang & Engelhard Jr, 2016) dan Bahasa Malaysia (bin Khairani & Razak, 2012). Hasil pengujian Rasch pada *TSES* versi Bahasa Malaysia menunjukan adanya bukti struktur instrumen dan reliabilitas alat ukur, yaitu skor *infit mean square* berada pada kisaran 0,75 - 1,25 dan nilai *Cronbach's Alpha* = 0,94., walau beberapa pernyataan pada skala tersebut dipersepsikan secara tidak konsisten oleh kelompok responden berbeda (*Differential Item Functioning* [*DIF*] rendah).

Pengujian *TSES* dengan pemodelan Rasch memungkinkan tersedianya informasi yang lebih akurat dan proses pengujian ini belum pernah dilakukan pada *TSES* versi Bahasa Indonesia. Dengan demikian, pemodelan Rasch yang dilakukan ini juga dapat mengatasi permasalahan perbedaan metrik ukur antar butir. Kalibrasi yang dibuat oleh pemodelan Rasch menempatkan butir serta subjek dalam ukuran yang sama (Engelhard, 2013; Boone et al., 2014). Skor yang dihasilkan bukan lagi skor mentah yang tidak diketahui *error* pengukurannya, melainkan skor pengukuran dengan satuan logit (*logarithm odd unit*) yang diketahui dengan pasti *error* pengukuran pada setiap butir (Sumintono & Widhiarso, 2013; 2015). Pengujian dengan menggunakan metode Rasch diharapkan dapat memberikan bukti empirik yang kuat terhadap penggunaan skala ini pada konteks yang lebih luas.

Penelitian ini bertujuan untuk melakukan evaluasi kualitas instrumen *TSES* dengan menggunakan analisis berbasis pemodelan Rasch yaitu *Rasch rating scale analysis* untuk memperoleh informasi psikometris tentang instrumen dari data skoring tiap butir *TSES*. Oleh karena itu, dengan pendekatan metode Rasch, ada tiga hipotesis penelitian.

Hipotesis 1: Skala *TSES* memiliki sifat unidimensional.

Hipotesis 2: Pilihan rating skala pada *TSES* berfungsi dengan efektif.

Hipotesis 3: Butir pada skala *TSES* dapat menggambarkan tingkat variasi responden secara efektif.

Metode

Sebagai langkah pertama, instrumen *TSES* yang berbahasa Inggris diterjemahkan ke dalam Bahasa Indonesia oleh peneliti dengan menggunakan prosedur *back-translation* (Brislin, 1970). Metode ini dilakukan dengan menterjemahkan instrumen Bahasa Inggris ke dalam Bahasa Indonesia, lalu instrumen diterjemahkan kembali ke dalam Bahasa Inggris untuk melihat konsistensi hasil terjemahan, yang melibatkan pihak Unit Pelayanan Bahasa, Universitas Negeri Jakarta (UNJ). Dalam proses *back-translation*, pakar bahasa menilai hasil penerjemahan kembali tersebut dan menemukan hasil penerjemahan kembali relatif sama dengan versi aslinya. Kisi-kisi (*blueprint*) *TSES* versi Bahasa Indonesia disediakan di Tabel 1.

Kode dimensi yang digunakan dalam proses analisis data adalah "E = *engagement*"; "I = *instructional*"; dan "C = *classroom management*". Misalnya untuk butir 1 kodenya E21, maksudnya E untuk *engagement*, 2 untuk indikator ke-2 (dapat mengatasi hambatan), dan 1 untuk butir ke-1 dalam urutan E2, demikian seterusnya. Skala peringkat (*rating scale*) yang digunakan adalah skor 1 sampai 9 seperti pada kuesioner asli dalam Bahasa Inggris. Intrumen juga dilengkapi dengan pertanyaan untuk variabel demografis yaitu, jenis kelamin, tingkat pendidikan terakhir dan status kepegawaian responden. Selanjutnya ins-

Tabel 1
Kisi-Kisi Instrumen TSES *Versi Bahasa Indonesia*

Faktor	Indikator	Butir	Total Butir
(a) Efikasi dalam meningkatkan keterlibatan atau motivasi siswa terhadap aktivitas pembelajaran *(student engagement)*	(1) Dapat menghadapi tantangan terkait dengan upaya meningkatkan keterlibatan anak terhadap pembelajaran di kelas.	2, 6, 9, 12, 22	8
	(2) Dapat mengatasi hambatan terkait dengan upaya meningkatkan keterlibatan anak terhadap pembelajaran di kelas	1, 4, 14	
(b) Efikasi dalam penyampaian materi ajar atau tugas alih pengetahuan *(instructional strategies)*	(1) Dapat menghadapi tantangan terkait dengan penyampaian materi ajar kepada anak	10, 11, 18, 20, 23, 24	8
	(2) Dapat mengatasi hambatan terkait dengan penyampaian materi ajar kepada anak	7, 17	
(c) Efikasi dalam pengelolaan kelas *(classroom management)*	(1) Dapat menghadapi tantangan terkait dengan tugas pengelolaan kelas	5, 8, 13, 16	8
	(2) Dapat mengatasi hambatan terkait dengan tugas pengelolaan kelas	3, 15, 19, 21	

trumen *TSES* berbahasa Indonesia disebarkan kepada responden. Kriteria responden penelitian ini yaitu individu dengan profesi guru pada tingkat sekolah dasar, tinggal dan mengajar di wilayah provinsi Daerah Khusus Ibukota (DKI) Jakarta. Kuesioner yang kembali dari responden sebanyak 255 buah dan semua data lengkap (tidak ada data hilang). Proses pemilahan data (*data filtering*) dilakukan dan didapati 10 responden yang bersifat *outlier* karena mengisi semua jawaban dengan skor 9 yang kemudian dihapus dari data, sehingga total data responden yang dianalisis adalah 245 orang. Informasi demografis responden terdapat pada Tabel 2.

Tahapan analisis data di awali dengan memasukkan data dari kuesioner dengan perangkat lunak *Microsoft Excel* (*MS Excel*), untuk disimpan menjadi berkas data dalam format .csv (*Comma Separated Value*) yang kompatibel untuk analisis Rasch. Setelah data dibuka dengan aplikasi *Winstep,* analisis psikometri *TSES* dapat dilakukan. Sebagaimana yang telah dijelaskan di atas, analisis Rasch menghasilkan konversi skor mentah butir *TSES* menjadi skor logit. Skor logit yang dihasilkan terdiri dari skor logit *item* (butir) dan skor logit *person* (responden). Skor ini digunakan dalam proses analisis selanjutnya dalam pemodelan Rasch.

Tahap analisis pertama adalah memastikan skala *TSES* memiliki sifat unidimensi. Untuk menentukan unidimensionalitas pada *TSES* penulis akan melihat hasil output dari pengolahan *Winstep* yaitu *Standardized Residual Variance* dengan pendekatan *Principal Component Analysis* (*PCA*). Hasil output

Tabel 2
Informasi Demografis Responden (N = 245)

Demografis	Jumlah	Persen (%)
Jenis Kelamin		
Laki-laki	71	71%
Perempuan	174	29%
Pendidikan terakhir		
Diploma 3 (D3)	15	6%
Diploma 4 (D4)	8	3%
Strata 1 (S1)	132	54%
Strata 2 (S2)	82	34%
Status Kepegawaian		
Calon Pegawai Negeri Sipil (CPNS)	10	4%
Pegawai Negeri Sipil (PNS)	162	66%
Honorer	73	30%

Catatan. *Status Kepegawaian. Guru yang digaji oleh Pemerintah (disebut "pegawai negeri") atau oleh sekolah/organisasi (disebut honorer).

akan menunjukkan nilai varians (*raw variance*), yaitu presentase minimal varians untuk dapat memenuhi standar unidimensionalitas adalah 20%. Tahap berikutnya adalah analisis peringkat skala. Analisis ini bertujuan untuk menguji validitas pilihan skala peringkat pada butir dengan tujuan melakukan konfirmasi apakah pilihan jawaban yang digunakan efektif dan tidak membingungkan responden. Nilai indeks *observed average* dan *Andrich threshold* digunakan sebagai acuan untuk menentukan keefektifan pilihan rating ini. Nilai disebut ideal jika skor kedua indeks mengalami peningkatan secara berurutan dari nilai negatif menuju positif. Hal ini menunjukan bahwa pilihan rating jawaban yang diberikan efektif, yaitu

dari skor logit rendah (negatif) menuju tinggi (positif). Ketidakberaturan pada skor pada kedua indeks menandakan adanya pilihan jawaban rating yang tidak efisien karena tidak banyak dipilih oleh responden.

Analisis berikut yang dilakukan adalah mengevaluasi kesesuaian butir (*item fit*). Analisis *item fit* memberikan informasi tentang tingkat kesukaran butir pada analisis Rasch. Hal ini bertujuan untuk melihat sejauh mana sebuah butir memiliki kecenderungan untuk mendapat jawaban setuju/*favorable* oleh responden. Analisis Rasch melalui aplikasi *Winstep* mampu menghasilkan peta *Wright* yang memberikan gambaran rata-rata logit butir dibandingkan dengan rata-rata logit responden. Analisis ini juga bisa dilihat pada indikator skor *Pearson measure* tiap butir pada output *Winstep* yang mengkaji *item measure*. Apabila nilai rata-rata lebih dari logit 0,00 menunjukan semakin banyak butir tersebut dijawab setuju oleh responden.

Pada bagian *item measure* tersebut dapat dilihat pula skor *Cronbach's Alpha* yang dapat dijadikan rujukan reliabilitas. Reliabilitas adalah ukuran yang menggambarkan keajegan sebuah butir ketika digunakan oleh responden yang berbeda dan pada waktu yang berbeda (Sumintono & Widhiarso, 2013). Reliabilitas pada pemodelan Rasch terdiri dari dua aspek, yaitu: (1) reliabilitas butir (*item*); dan (2) reliabilitas responden (*person*), dan skor reliabilitas yang di harapkan adalah > 0,08 (Sumintono & Widhiarso, 2013). Uji kualitas butir dilakukan dengan cara mengacu pada tabel *Item Fit Order*. Pada tabel ini di informasikan urutan butir dari yang paling tidak sesuai dengan data (*item misfit*). Menurut Boone, Staver, dan Yale (2014), hal ini ditentukan oleh tiga indeks yaitu: (1) *Outfit Mean-Square (MNSQ;* dalam rentang 0,5 sampai 1,5); (2) *Outfit z-standarized (ZSTD;* dalam rentang - 2 sampai + 2); dan (3) *point measure correlation* (0,4 sampai 0,85). Apabila nilai tersebut tidak tercapai, butir tersebut dapat dikategorikan tidak *fit* (*misfit*).

Terakhir, analisis bias pada butir dilakukan dengan mengacu pada fungsi *DIF*. *DIF* pada umumnya berfungsi untuk menentukan kesulitan butir. Namun, *DIF* juga berfungsi untuk menilik apakah butir-butir mempunyai bias dalam kategori responden tertentu atau tidak (Sumintono & Widhiarso, 2013). Butir dengan nilai probabilitas dibawah 5% (0,5) maka akan dikategorikan sebagai bias.

Tabel 3
Reliabilitas Persons dan Item Instrumen TSES

	Rata-rata Logit (SD)	*Separasi*	*Reliabilitas*	*Cronbach's Alpha*
Person	2,22 (1,46)	4,03	0,94	0,95
Item	0,00 (0,27)	2,72	0,88	

Hasil

Berikut ini akan dipaparkan mengenai properti psikometris yang dihasilkan dari pengolahan data menggunakan pemodelan Rasch, yang meliputi reliabilitas dan validitas dari instrumen *TSES* yang didapatkan dari 245 responden yang berpartisipasi dalam penelitian ini. Secara umum, nilai rata-rata logit responden menunjukan kecenderungan yang tinggi (+ 2,22 logit), ini menunjukkan bahwa keseluruhan responden mempunyai skor *teacher efficacy* yang positif (lebih tinggi dari nilai logit 0,00 yang merupakan standar logit tingkat kesulitan butir). Hal ini makin jelas dengan nilai standar deviasi-sebesar 1,46 logit, yang makin menunjukkan kisaran logit dari sebagian besar responden (68%) yang berada di *range* positif (lebih dari 0,00 logit). Nilai rata-rata logit butir sebesar 0,00 logit yang menunjukkan tingkat kesulitan butir standar, namun nilai deviasi standarnya adalah 0,27 yang menunjukkan kurang beragamnya tingkat kesulitan butir-butir dari instrumen *TSES* ini (lihat pada Tabel 3).

Bukti Reliabilitas Instrumen

Dalam konteks pemodelan Rasch, indeks yang digu-nakan untuk mengukur keajegan adalah reliabilitas dan separasi logit butir dan person, seperti ditunjukkan pada Tabel 3. Baik reliabilitas responden (*person)* maupun butir (*item*) menunjukkan nilai yang tinggi (berturut-turut 0,94 dan 0,88) demikian juga dengan nilai indeks *Cronbach's Alpha*. Konsistensi hasil analisis menggunakan dua metode ini menunjukkan/ dapat diinterpretasi sebagai butir-butir pada *TSES* memiliki konsistensi ketika di isi oleh para responden penelitian.

Estimasi parameter separasi butir memiliki nilai 4,03 yang berarti butir tersebut dapat mengelompokkan responden berdasarkan tingkat efikasi-nya, sementara estimasi parameter separasi person memiliki nilai 2,72 (dibulatkan menjadi 3,00) yang berarti terdapat tiga kategori kelompok responden berdasarkan

urutan skor efikasinya. Dengan nilai separasi responden adalah 4, maka instrumen yang digunakan bisa memetakan atau mengelompokkan responden. Ini menunjukkan bahwa responden yang berpartisipasi dalam riset tidak homogen melainkan heterogen dan menunjukkan responden yang diperoleh merupakan representasi populasi. Info reliabilitas menunjukkan bahwa walaupun aspek *item* dan *person* memang baik, namun yang person menunjukkan indeks sedikit lebih baik.

Bukti Struktur Penyusun Instrumen

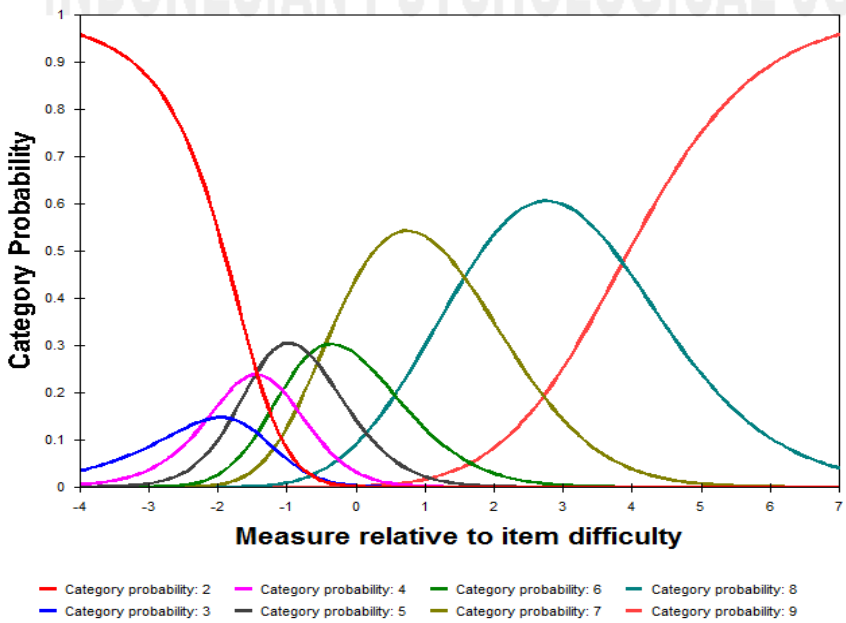
Tahapan selanjutnya pada pemodelan Rasch di penelitian ini mencakup argumen dalam hal unidimensionalitas, kesesuaian skala peringkat yang digunakan, indeks ketepatan butir-model (*item fit statistik*), serta kemungkinan adanya butir yang bias (*DIF*). Berikut ini akan disajikan hasil dari setiap aspek validitas di atas.

Unidimensionalitas skala. Hasil pengujian *Standardized Residual Variance* pada instrumen TSES menunjukkan bahwa *raw variance* bernilai 49,4% dan dapat dikategorikan memuaskan (Fisher, 2007). Hal ini menunjukkan persentase minimum varians sebesar 20% terpenuhi. Selain itu, pada *unexplained variance* berkisar antara 3,0% hingga 4,6%, dimana tidak ada yang melebihi 15%, yang menunjukkan bahwa instrumen memang mengukur satu variabel saja yaitu *TE*.

Tabel 4
Statistik Analisis Skala Peringkat

Rating (skor)	*Observed average*	*Andrich Treshold*
2	- 0,43	None
3	0,61	- 0,70
4	- 0,37	- 2,16
5	0,22	- 1,45
6	0,38	- 0,68
7	1,48	- 0,45
8	2,45	1,57
9	3,86	3,86

Kesesuaian skala peringkat (Rating Scale). Hasil analisis skala peringkat ditunjukkan pada Tabel 4. Utilitas skala peringkat menunjukkan hanya delapan *rating* yang digunakan (skor 2 ke 9), tidak ada satu pun responden yang memilih skor 1 (*tidak mampu*); hal ini menunjukkan secara empirik skala peringkat yang diberikan seharusnya lebih sedikit. Pada indeks *observed average* terlihat nilai yang naik turun (- 0,43 kemudian + 0,61 dan turun ke - 0,37) pada rating 2, 3 dan 4. Hal ini menunjukkan responden tidak terlalu pasti dengan pilihan peringkat yang diberikan, yaitu pada skor 2, 3 dan 4. Demikian juga pada indikasi *Andrich threshold* yang bergerak pada - 0,70 kemudian turun ke - 2,16 dan lalu naik ke - 1,45 pada skor 3, 4, dan 5. Kedua hal ini menunjukkan bahwa responden tidak begitu paham dengan banyaknya pilihan rating yang diberikan, khususnya antara



Gambar 1. Analisis Skala Peringkat *TSES*.

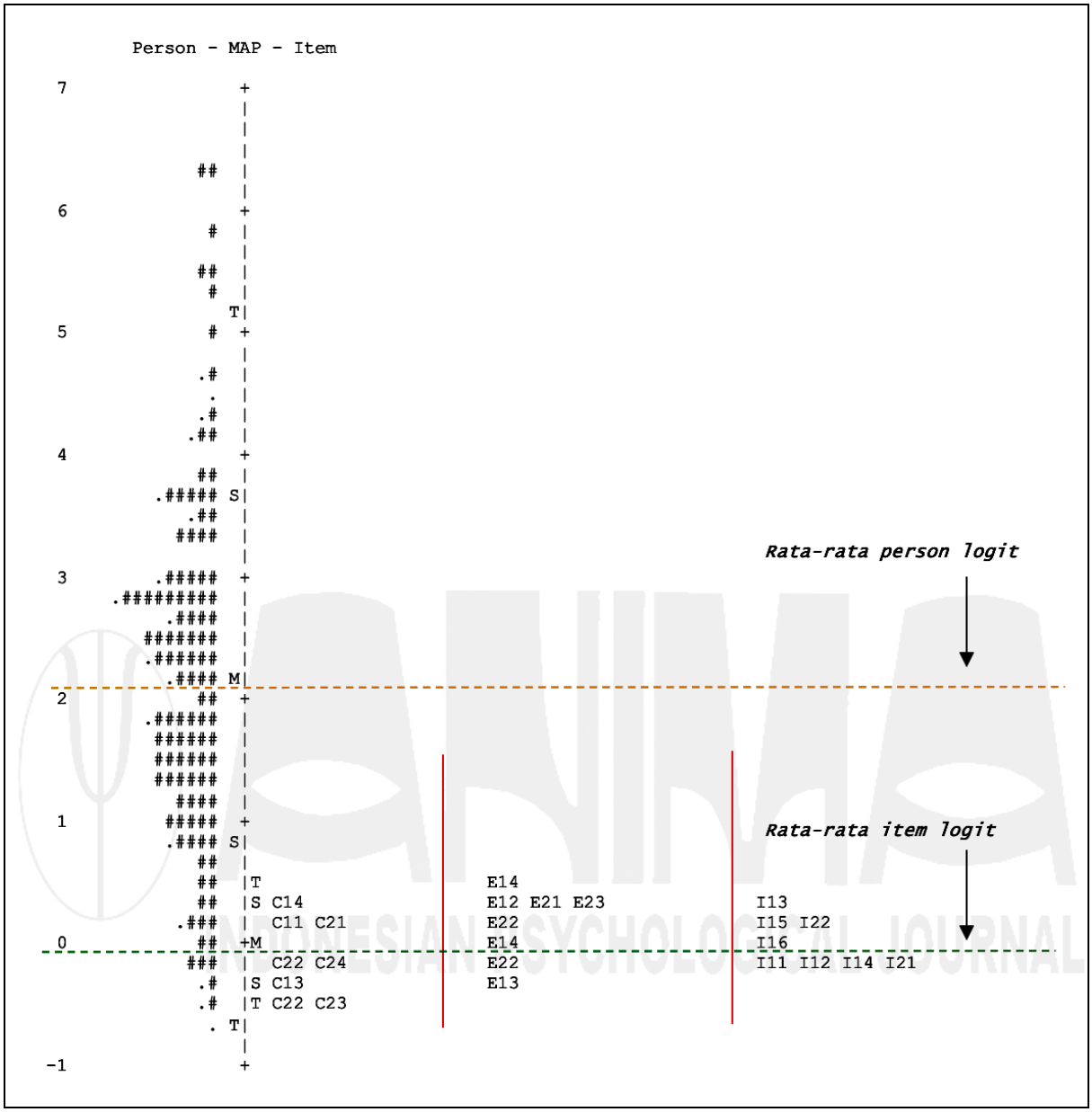
Tabel 5
Statistik Kesesuaian Butir (Item Fit Statistic)

No	Butir	*Logit*	*Standard Error Measurement*	*Outfit MNSQ*	*Outfit ZSTD*	*Point Measure Correlation*
1	E21	+ 0,39	0,08	1,07	+ 0,70	0,71
2	E12	+ 0,36	0,08	0,82	- 1,86	0,71
3	C21	+ 0,18	0,09	1,14	+ 1,35	0,66
4	E22	+ 0,18	0,09	1,44	+ 3,95	0,61
5	C11	+ 0,14	0,09	1,04	+ 0,43	0,68
6	E22	- 0,18	0,09	0,98	- 0,12	0,65
7	I21	- 0,15	0,09	1,12	+ 1,15	0,64
8	C22	- 0,12	0,09	0,95	- 0,45	0,67
9	E13	- 0,30	0,09	1,07	+ 0,67	0,61
10	I11	- 0,20	0,09	0,99	- 0,06	0,63
11	I12	- 0,15	0,09	0,97	- 0,24	0,62
12	E14	- 0,05	0,09	0,88	- 1,24	0,66
13	C13	- 0,39	0,10	1,12	+ 1,15	0,66
14	E23	+ 0,31	0,08	0,72	- 3,07	0,75
15	C22	- 0,48	0,10	0,98	- 0,16	0,70
16	C14	+ 0,25	0,09	0,76	- 2,63	0,74
17	I22	+ 0,21	0,09	0,86	- 1,39	0,70
18	I13	+ 0,30	0,08	0,82	- 1,93	0,73
19	C23	- 0,43	0,10	0,97	- 0,26	0,66
20	I14	- 0,18	0,09	0,70	- 3,35	0,75
21	C24	- 0,19	0,09	0,99	- 0,06	0,69
22	E14	+ 0,43	0,08	1,28	+ 2,61	0,59
23	I15	+ 0,12	0,09	0,97	- 0,27	0,64
24	I16	- 0,05	0,09	1,02	+ 0,27	0,67

skala peringkat dengan skor 2 sampai 5.Hal yang berbeda didapati pada skala peringkat dengan skor yang lebih tinggi, yaitu skor 6, 7, 8 dan 9. Baik indeks *observed average* maupun *Andrich threshold* menunjukkan peningkatan nilai indeks, ini mengindikasikan bahwa skala peringkat dalam pandangan responden memang terpisah dan bisa dibedakan dengan mudah. Grafik skala peringkat dalam Gambar 1. mendukung dugaan ini karena kurva probabilitas memilih satu kategori pada kategori rendah terlihat saling tumpang tindih tanpa pemisahan yang dapat teramati jelas. Hal ini menunjukkan pemisahan *rating* yang susah dibedakan oleh responden dibanding pilihan *rating* pada skor yang besar.

Item Fit. Hasil analisis kesesuaian butir ditampilkan pada Tabel 5, yang diurutkan berdasar nomor butir. Nilai logit butir menunjukkan tidak terdapat butir dengan tingkat kesulitan terlalu tinggi (mudah untuk diberikan jawaban setuju oleh responden). Kisaran tingkat kesukaran butir, atau butir yang akan dijawab dengan tidak setuju, berada antara -0,48 logit (Butir 15 atau C22) sampai ke + 0,43 logit (Butir 22 atau E14), dimana jumlah responden yang berada dalam rentang ini sejumlah 32 orang saja (13%) (lihat pada Gambar 2). Selanjutnya, untuk mengetahui kesesuaian antara data butir dengan model adalah dengan mengacu pada tabel *item fit statistic* (Tabel 5). Bila dilihat pada Tabel 5, semua butir memenuhi kriteria *fit statistic* untuk *Outfit MNSQ*; sedangkan pada kriteria *Outfit ZSTD* terdapat lima butir yang diluar batas ketepatan (yaitu E22, E14, C14, I14 dan E23). Dari aspek *point measure correlation* semua butir yang ada memenuhi syarat berada dalam kisaran ketepatan butir-model; dengan semua nilai yang positif hal ini menunjukkan tidak adanya polaritas butir dalam instrumen *TSES* ini. Secara keseluruhan tidak ada butir yang tidak memenuhi ketiga syarat dalam aspek validitas kesesuaian statistik.

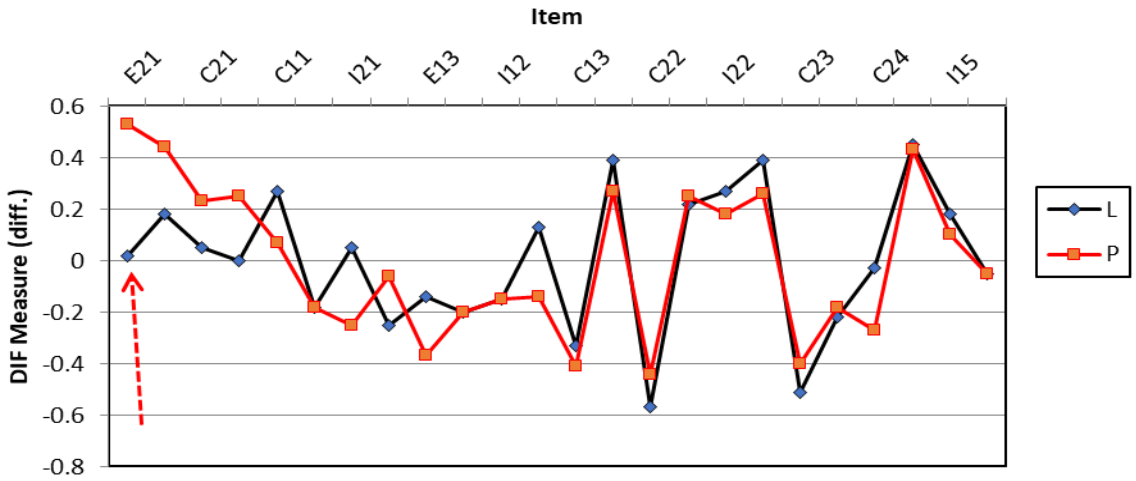
Peta *Wright* (*item-person map*) pada Gambar 2 membantu untuk menjelaskan kesesuaian antara *item* (butir) dengan *person* (responden). Peta *Wright* menunjukkan ketiga dimensi *TE* dalam instrumen *TSES* berada dalam rentang yang tidak terlalu berjauhan dan berada dalam tingkatan yang mudah disetujui oleh responden. Dalam peta tersebut dapat



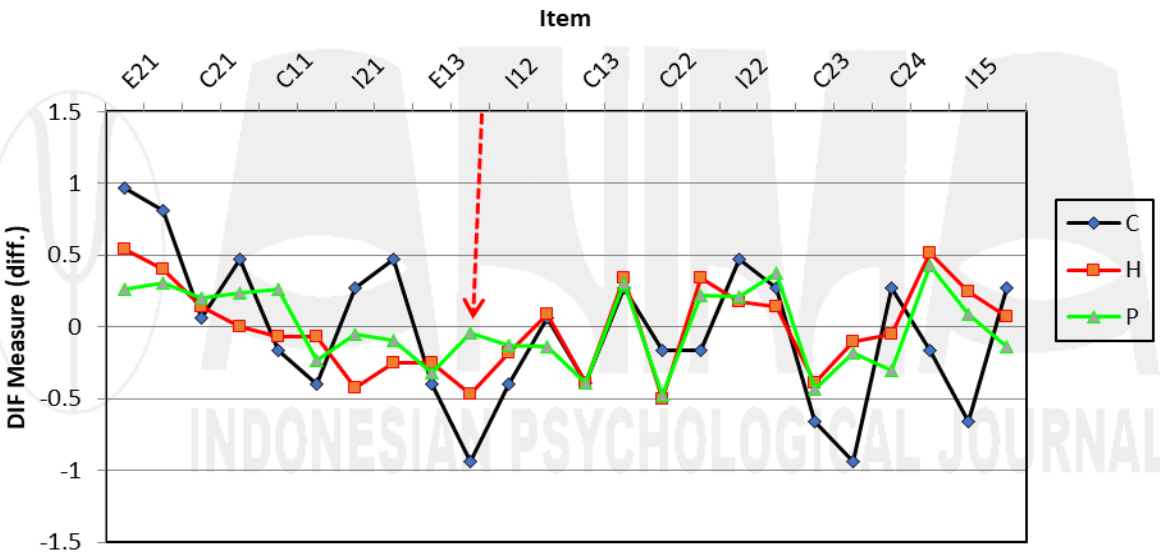
Gambar 2. Peta *Wright* Instrumen *TSES*.
Keterangan. C = Classroom Management; *E = Instructional Engagement*; *I = Instructional Strategist.*

terlihat beberapa responden memiliki nilai ekstrim, yaitu pada bagian sisi sebelah kiri peta, yang berlokasi diatas tanda T (maksudnya adalah dua kali deviasi standar dari rata-rata) atau di atas nilai logit + 5,00 logit. Pada aspek butir (bagian kanan peta), terlihat bahwa C11 dan C21 berada dalam kisaran nilai logit yang sama;demikian juga dengan C22 dengan C24. Pada dimensi *instructional engagement* Butir E12, E21 dan E23 juga berada dalam kisaran nilai logit yang sama. Demikian juga halnya dengan rentang nilai logit Butir I11, I12 dan I14 pada dimensi *instructional strategist* berada dalam baris yang sama.

Analisis DIF. Penulis juga melakukan analisis untuk mengidentifikasi butir yang memiliki fungsi berbeda antar kelompok responden atau dikenal sebagai butir dalam kategori *DIF*. Terdapat tiga varians data variabel demografis yang digunakan untuk mengetahui adanya butir yang bias (jenis kelamin, pendidikan, dan status kepegawaian). Hasil analisis *DIF* menunjukan terdapat dua butir yang dianggap bias ketika dibandingkan dengan variabel demografis



Gambar 3. Analisis *DIF* berdasar variabel jenis kelamin.
Keterangan. L = Laki-Laki; P = Perempuan.



Gambar 4. Analisis *DIF* berdasar variabel status kepegawaian.
Keterangan. C = CPNS; H = Honorer; P = PNS

jenis kelamin, yaitu pada Butir E21 (nilai probabilitas 0,0084). Sementara ketika dibandingkan berdasarkan varians demografis status kepegawaian, butir I11 (nilai probabilitas 0,0427) juga tampak bias. Tidak terdapat butir yang bias berdasar pola respon dari variabel demografis tingkat pendidikan terakhir.

Penjelasan lebih lanjut tentang analisis *DIF* dapat dilihat di Gambar 3 dan Gambar 4. Pada Gambar 3 yang menunjukkan perbedaan pola respon pada butir berdasar variabel jenis kelamin, kurva warna hitam melambangkan guru lelaki dan warna merah guru perempuan. Bila kurva mendekati batas atas artinya butir tersebut mempunyai tingkat kesulitan tinggi, sedangkan bila kurva mendekati batas bawah artinya tingkat kesulitannya rendah. Untuk Butir E21 (butir yang mengalami *DIF*) pada Gambar 3 menunjukkan perbedaan pola respon yang kontras (tanda panah), butir dianggap mudah disetujui oleh guru lelaki dan saat yang sama dianggap susah disetujui oleh guru perempuan. Bandingkan dengan butir yang paling ujung yaitu I15, tidak terdapat perbedaan pola respon antara guru lelaki dan perempuan.

Pada Gambar 4 yang menampilkan analisis *DIF* berdasar status kepegawaian yang terdiri dari tiga kelompok. Butir I11 yang mengalami *DIF* (tanda panah) menunjukkan pola respon yang berbeda bagi

tiga kelompok ini, bahwa kelompok guru CPNS (Calon Pegawai Negeri Sipil; kurva hitam) menganggap butir ini mudah disetujui dibandingkan kelompok guru honor (kurva warna merah), namun butir ini lebih dianggap paling susah disetujui oleh kelompok guru PNS (Pegawai Negeri Sipil; kurva warna hijau).

Diskusi

Berdasarkan hasil analisis data dengan pemodelan Rasch terlihat bahwa keseluruhan responden mempunyai tingkat *TE* yang positif atau tinggi (rata-rata + 2.22 logit). Hasil analisis juga menunjukkan bahwa instrumen yang digunakan bisa memetakan atau mengelompokkan responden. Hasil pemetaannya adalah responden tidak homogen namun heterogen. Hasil ini menunjukkan responden yang dilibatkan dalam penelitian merupakan representasi populasi yang reliabel. Selain itu, koefisien reliabilitas person menunjukkan tingkat reliabilitas responden yang baik sekali.

Sementara itu dari sisi kualitas butir, hasil analisis Rasch mengindikasikan tingkat kesulitan butir cenderung kurang bervariasi. Artinya sebagian besar butir lebih mudah dijawab setuju oleh responden, sehingga tidak memiliki variasi jawaban. Hal ini memungkinkan terjadinya *social desirability*. Dua puluh empat butir yang digunakan cenderung dijawab secara *favorable* bagi responden sehingga fungsi butir untuk mengukur *TE* perlu ditinjau ulang; secara umum hanya 13% (32 responden) saja yang secara efektif bisa terukur berdasar kemungkinan butir dijawab secara setuju atau tidak setuju oleh responden.

Hipotesis pertama penelitian ini terdukung bahwa *TSES* memiliki sifat unidimensional. Berdasarkan kajian unidimensionalitas instrumen dan ketepatan butir-model (*fit statistic*), *TSES* memang hanya mengukur satu variabel saja yaitu *TE* dan semua butir memenuhi syarat kesesuaian statistik serta semua butir berfungsi secara seragam. Semua butir *TSES* tidak mengalami polaritas (nilai *point measure correlation* semunya positif).

Adanya beberapa butir pada dimensi yang sama dengan kisaran nilai logit yang tidak jauh berbeda menunjukkan bahwa instrumen sebenarnya dapat lebih disederhanakan dengan hanya memilih salah satu dari beberapa butir yang berfungsi sama tersebut. Butir dengan nilai logit yang sama menunjukkan butir itu mengukur konsep yang sama (Bond & Fox, 2015; Boone et al., 2014). Sebagai contoh pada Butir 11 ("*Sejauh mana anda dapat membuat pertanyaan yang bagus pada siswa*?" dan 12 ("*Seberapa besar usaha anda mengembangkan kreatifitas siswa*?"). Kedua butir ini mengukur dimensi *instructional strategy*. Berdasarkan hasil analisa, kedua pertanyaan ini tidak terlalu efisien karena mengukur konsep yang sama; pun secara kualitatif, dua butir ini memang terkesan tumpang tindih (*overlap*), dengan demikian revisi atas butir-butir ini yaitu cukup dipilih salah satu saja.

Hal ini mengindikasikan bahwa perombakan butir-butir perlu dilakukan di semua dimensi *TE* dalam instrumen *TSES*. Cara yang bisa dilakukan untuk memperbaiki instrumen ini adalah mengubah atau merevisi butir menjadi lebih susah untuk lebih mudah disetujui oleh responden (*agreeable*) khususnya apabila nilai logit dalam satu dimensi tidak jauh beda. Cara lain yang bisa digunakan adalah mengubah beberapa butir menjadi *unfavorable statement* (*negative statement*) yang diasumsikan akan membuat responden berpikir berbeda dan lebih serius dibandingkan jika semua butirnya *favorable* (positif). Walaupun demikian dari segi pengacakan butir sudah bekerja dengan baik.

Hipotesis kedua yang berkaitan dengan tingkat variasi respon yang dikaji berdasarkan pilihan rating skala belum sepenuhnya terdukung. Dalam analisis Rasch, pilihan jawaban dalam skala Likert (*rating scale*) dievaluasi efektivitasnya. Pada *TSES*, walaupun diberikan skala peringkat 1 sampai 9 namun kenyataannya responden hanya memilih dalam kisaran skor 2 - 9, artinya *rating* empirik bagi responden yang berfungsi efektif hanya delapan *rating* saja. Saat yang sama dengan skala peringkat delapan pun, pilihan jawaban responden untuk dapat dikategorikan *rating* rendah (skor 2 - 5) juga tidak berjalan efektif. Dengan demikian, hal ini menunjukkan pilihan *rating* yang diberikan terlalu panjang *range*-nya (1 - 9), dan kondisi ini potensial membingungkan responden. Solusinya adalah penyederhanaan rating, seperti 1 - 4 atau 1 - 5, dengan setiap pilihan rating diberi penjelasan yang memadai seperti halnya skala peringkat Likert (Bond & Fox, 2015). Kondisi ini dapat disebabkan gaya respon yang berbeda pada setiap individu yang dipengaruhi oleh perasaan individu tentang wadah merespon yang disediakan dalam instrumen. Ada individu yang merasa responnya terwadahi dengan hanya dua pilihan jawaban ("*Ya*" dan "*Tidak*"), namun ada juga individu yang merasa lebih mudah merespon jika pilihannya dengan opsi berjenjang. Preferensi berbeda terhadap jumlah respon seperti itu dapat menghasilkan perbedaan struktural yang menurunkan kualitas informasi yang diperoleh

dari pengisian instrumen dalam suatu survei (Widhiarso, 2016).

Hipotesis terakhir studi ini tentang tingkat variasi respon berdasarkan karakteristik partisipan terbukti efektif. Temuan ini diperoleh dari analisis *DIF*. Penulis melakukan memberikan rekomendasi perbaikan butir yang menunjukkan indikasi bias berdasarkan pada analisis *DIF*, yaitu Butir E21 ("*Seberapa baik usaha Anda merespon siswa yang suka memberontak*") dan I15 ("*Seberapa besar usaha Anda menenangkan siswa yang mengganggu dan berisik*?"). Kedua butir ini perlu direvisi supaya menghindarkan responden dari menjawab secara berbeda untuk kategori demografis tertentu (Boone et al., 2014). Secara kualitatif, kedua butir ini memang berpotensi dapat direspon berbeda. Hal ini mungkin dapat disebabkan persepsi terhadap perbedaan derajat otoritas guru berdasarkan status kepegawaian mereka dalam mempertimbangkan kemungkinan menyetujui atau tidak menyetujui butir-butir tersebut. Perlu pendalaman lebih lanjut terhadap asumsi tersebut. Walau demikian, dengan hanya dua butir saja yang terindentifikasi bias, menunjukkan kualitas butir secara keseluruhan tidak terlalu bermasalah ketika digunakan pada kategori sampel yang berbeda.

Terlepas dari semua temuan penelitian di atas, studi ini memiliki beberapa catatan keterbatasan yang dapat diperbaiki di masa yang akan datang. Keterbatasan pertama yaitu jumlah sampel penelitian yang relatif kecil bila dibandingkan dengan populasi sampel. Penentuan ukuran sampel yang sesuai dapat membantu untuk mengetahui varians butir berdasarkan karakteristik kelompok yang lebih beragam. Menurut Herrera dan Gómez (2008), keterbatasan sampel dalam uji ini khususnya terkait dengan interpretasi skor *DIF* dalam mendeteksi apakah terjadi bias pada butir. Ketidak seimbangan jumlah sampel pada masing-masing sub kelompok akan memengaruhi derajat keakuratan *DIF* dalam mendeteksi bias pada butir (Herrera & Gómez, 2008) Keterbatasan berikut nya adalah proses adaptasi instrumen dari versi Bahasa Inggris yang diterjemahkan oleh penulis sendiri, hanya berfokus pada tata Bahasa saja, dan tidak terlalu mengkaji detil konteks kesesuaian butir dengan karakteristik responden yang beragam. Menurut Cha et al. (2007), penting untuk melibatkan beberapa orang dari kelompok karakteristik partisipan dalam proses penerjemahan, sehingga didapatkan hasil terjemahan yang akurat. Berdasarkan keterbatasan-keterbatasan tersebut, maka penting untuk melakukan analisa lanjutan, baik itu dengan menggunakan pendekatan metode kuantiatif, atau kualitatif untuk menguji kualitas *TSES* versi Bahasa Indonesia dengan tujuan memverifikasi hasil penelitian yang sudah dilakukan.

Simpulan

Berdasarkan hasil analisis data menggunakan pemodelan Rasch, kualitas instrumen *TSES* versi Bahasa Indonesia pada beberapa hal sudah cukup baik. Instrumen ini memenuhi kaidah dimensionalitas yaitu hanya mengukur tentang variabel *TE* dan hanya memiliki dua butir yang bias serta 24 butirnya tidak mengalami polaritas. Di sisi lain terdapat beberapa kekurangannya yang dapat diperbaiki untuk penggunaannya di masa depan, yaitu tingkat kesulitan butir yang rendah sehingga kurang dapat mengungkap *TE* pada responden yang nilai logit-nya tinggi (responden yang cenderung selalu memberikan jawaban setuju).

Beberapa hal yang perlu dilakukan terkait perbaikan instrumen ke depannya adalah dengan menyederhanakan skala peringkat (rentang pilihan interval) menjadi lebih pendek, merevisi berbagai butir dalam satu dimensi menjadi butir dengan tingkat kesulitan lebih tinggi serta memodifikasi format beberapa butir ke dalam bentuk *unfavorable*. Penelitian selanjutnya juga dapat memperbaiki penerjemahan instrumen ke Bahasa Indonesia yang dapat memengaruhi pemahaman partisipan dan respon partisipan terhadap butir-butir *TSES*. Penerjemahan *TSES* dapat melibatkan pakar bahasa serta individu yang berasal dari kelompok karakteristik partisipan.

References

Andrich, D. (1988). *Rasch model for measurement* (*Series: Quantitative application in the social sciences*). SAGE Publications. https://doi.org/10.4135/9781412985598

Bandura, A. (1994). Self-efficacy. In R. J. Corsini (Ed.), *Encyclopedia of psychology* (2nd ed.) (vol. 3. pp. 368-369). John Wiley & Sons, Inc.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

bin Khairani, A. Z., & Razak, N. b. A. (2012). An analysis of the teachers' sense of efficacy scale within the Malaysian context using the Rasch measurement model. *Procedia - Social and Behavioral Sciences*, *69*(December), 2137-2142. https://doi.org/10.1016/j.sbspro.2012.12.178

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185-216. https://doi.org/10.1177/135910457000100301

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://www.routledge.com/Applying-the-Rasch-Model-Fundamental-Measurement-in-the-Human-Sciences/Bond/p/book/9780415833424

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://www.springer.com/gp/book/9789400768567

Cha, E.-S., Kim, K. H., & Erlen, J. A. (2007). Translation of scales in cross-cultural research: Issues and techniques. *Journal of Advanced Nursing*, *58*(4), 386-395. https://doi.org/10.1111/j.1365-2648.2007.04242.x

Chang, M.-L., & Engelhard, G., Jr. (2016). Examining the Teachers' Sense of Efficacy Scale at the item level with Rasch measurement model. *Journal of Psychoeducational Assessment*, *34*(2), 177-191. https://doi.org/10.1177/0734282915593835

Darling-Hammond, L. (2003). Keeping good teachers: Why it matters and what leaders can do. *Educational Leadership: Journal of the Department of Supervision and Curriculum Development, N. E. A.*, *60*(8), 6-13. https://www.researchgate.net/publication/242663183_Keeping_Good_Teachers_Why_It_Matters_What_Leaders_Can_Do

Duffin, L. C., French, B. F., & Patrick, H. (2012). The Teachers' Sense of Efficacy Scale: Confirming the factor structure with beginning pre-service teachers. *Teaching and Teacher Education*, *28*(6), 827-834. https://doi.org/10.1016/j.tate.2012.03.004

Dixon, F. A., Yssel, N., McConnell, J. M., & Hardin, T. (2014). Differentiated instruction, professional development, and teacher efficacy. *Journal for the Education of the Gifted*, *37*(2),111-127. https://doi.org/10.1177/0162353214529042

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge. https://www.routledge.com/Invariant-Measurement-Using-Rasch-Models-in-the-Social-Behavioral-and/Engelhard-Jr/p/book/9780415871259

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095. https://www.rasch.org/rmt/rmt211m.htm

Herrera, A-N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity*, *42*(6), 739. https://doi.org/10.1007/s11135-006-9065-z

Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, *102*(3), 741-756. https://doi.org/10.1037/a0019237

Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998-2009: Signs of progress or unfulfilled promise? *Educational Psychology Review*, *23*(1), 21-43. https://doi.org/10.1007/s10648-010-9141-8

Kleinsasser, R. C. (2014). Teacher efficacy in teaching and teacher education. *Teaching and Teacher Education*, *44*(November), 168-179. https://doi.org/10.1016/j.tate.2014.07.007

Maddux, J. E., & Lewis, J. (1995). Self-efficacy and adjustment: Basic principles and issues. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 37-68). Springer. https://link.springer.com/chapter/10.1007/978-1-4419-6868-5_2

Mehdinezhad, V., & Mansouri, M. (2016). School principals' leadership behaviours and its relation with teachers' sense of self-efficacy. *International Journal of Instruction*, *9*(2), 51-60. https://doi.org/10.12973/iji.2016.924a

Nie, Y., Lau, S., & Liau, A. (2012). The teacher efficacy scale: A reliability and validity study. *The Asia-*

Pacific Education Researcher, *21*(2), 414-421. http://hdl.handle.net/10497/14287

Scherer, R., Jansen, M., Nilsen, T., Areepattamannil, S., & Marsh, H. W. (2016). The quest for comparability: Studying the invariance of the Teachers' Sense of Efficacy Scale (TSES) measure across countries. *PLoS ONE*, *11*(3), e0150829. https://doi.org/10.1371/journal.pone.0150829

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial* [Application of Rasch model for research in social sciences]. TrimKom Publishing House. https://www.researchgate.net/publication/256498376_Aplikasi_Model_Rasch_untuk_Penelitian_Ilmu-Ilmu_Sosial

Sumintono, B. & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* [Application of Rasch modelling on educational assessment]. TrimKom Publishing House. https://www.researchgate.net/publication/282673464_Aplikasi_Pemodelan_Rasch_pada_Assessment_Pendidikan

Soodak, L. C., & Podell, D. M. (1996). Teacher efficacy: Toward the understanding of a multi-faceted construct. *Teaching and Teacher Education*, *12*(4), 401-411. https://doi.org/10.1016/0742-051X(95)00047-N

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783-805. https://doi.org/10.1016/S0742-051X(01)00036-1

Tschannen-Moran, M., & Gareis, C. R. (2004). Principals' sense of efficacy: Assessing a promising construct. *Journal of Educational Administration*, *42*(5), 573-585. https://doi.org/10.1108/09578230410554070

Tucker, C. M., Porter, T., Reinke, W. M., Herman, K. C., Ivery, P. D., Mack, C. E., & Jackson, E. S. (2005). Promoting teacher efficacy for working with culturally diverse students. *Preventing School Failure: Alternative Education for Children and Youth*, *50*(1), 29-34. https://doi.org/10.3200/PSFL.50.1.29-34

Ware, H., & Kitsantas, A. (2007). Teacher and collective efficacy beliefs as predictors of professional commitment. *The Journal of Educational Research*, *100*(5), 303-310. https://doi.org/10.3200/JOER.100.5.303-310

Widhiarso, W. (2016). Eksplorasi gaya respon ekstrem dalam mengisi kuesioner [Exploration of extreme response style in filling in questionnaires]. *Jurnal Psikologi*, *43*(1), 16-29. https://doi.org/10.22146/jpsi.8703

Wilcox-Herzog, A., & Ward, S. L. (2004). Measuring teachers' perceived interactions with children: A tool for assessing beliefs and intentions. *Early Childhood Research & Practice*, *6*(2). https://ecrp.illinois.edu/v6n2/herzog.html

Yeo, L. S., Ang, R. P., Chong, W. H., Huan, V. S., & Quek, C. L. (2008). Teacher efficacy in the context of teaching low achieving students. *Current Psychology*, *27*(3), 192-204. https://doi.org/10.1007/s12144-008-9034-x