Kasetsart Journal of Social Sciences 42 (2021) 599–606

Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

The estimation of test item difficulty using focus group discussion approach on the semantic differential scale

Rukli Rukli*, Ma'rup Ma'rup, Erni Ekafitria Bahar, Rezki Ramdani

Mathematics Education Study Program, Faculty of Teacher Training and Education, Muhammadiyah University Makassar, Makassar, South Sulawesi 90221, Indonesia

Article Info

Article history:
Received 20 April 2020
Revised 16 October 2020
Accepted 29 October 2020
Available online 31 July 2021

Keywords:
CTT,
estimation,
focus group discussion,
Semantic Differential Scale,
test item difficulty

Abstract

This study aimed to estimate the test item difficulty by comparing the approach to the semantic differential scale and the item and test analysis program. The research used a quantitative-comparative approach, involving 35 samples of 14 teachers and 21 students. Four group pairs were compared. The comparative used the Kruskal-Wallis test then used the Mann Whitney test, with the .01 significance level. The results showed as follows. First, students' involvement in assessing the test item difficulty is not just providing true or false information. However, the teacher can provide feedback to diagnose student difficulties with the material. Second, student groups are an alternative approach to estimating the difficulty of test items in schools as an option to replace the program. Third, teacher-student groups are an alternative approach to estimating the difficulty of test items in schools as an option to replace the program. The combination of teacher and student was estimated difficulty of the closest item test to the program's output. It refers to the approach of assessment for learning.

Introduction

Test items on multiple-choice tests have several characteristics to suit the scope of the theory test's approach. The Classical Test Theory (CTT) approach, multiple-choice items, has three characteristics: difficulty, discrimination power, and distractor efficiency. It is different from Item Response Theory (IRT) regarding assumptions and their implementation (Raykov & Marcoulides, 2016). These three characteristics determine whether a test item is right or not suitable to be tested on students. However, related to the continuous line of the examinees' ability, the difficulty has particular important virtues (Jabrayilov, Emons, & Sijtsma, 2016).

The difficulty of the test items relates to the continuous line of abilities of the examinee group. Besides that, the difficulty is empirically most easily determined. However, teachers sometimes do not do this so that the test items are sometimes not suitable for use as a measurement tool (Boopathiraj & Chellamani, 2013). The task is the teacher's obligation as part of the learning process. Nevertheless, for teachers, assessing test item difficulty will add to other assignment burdens, so there needs to be another way without consuming time, energy and thought. Likewise,

* Corresponding author.
E-mail address: rukli@unismuh.ac.id (Rukli).

https://doi.org/10.34044/j.kjss.2021.42.3.22
2542-3151/© 2021 Kasetsart University.

the teacher does not need to test the test items in the field, so there is no need to use a computer application.

Usually, calculating the difficulty of test items uses a computer program (Martinková et al., 2017). For example, the Program of Item and Test Analysis. The program uses a CTT approach (Juškaite, 2018). Although the program is easy to apply at the class level, it still requires a different understanding of some of the other applications when entering data into the program, such as excel, word, and notepad+++. Similarly, the program requires certain specifications in terms of computer operating systems. In the need for additional software to be compatible, it is necessary to use another simple approach for teachers in the classroom, but it can be a solution to determine the test items' difficulty.

Several approaches can be taken to solve a problem other than the statistical or computer program approach, such as the adjustment approach. The adjustment approach is an approach to solve a specific item with the minimum requirements (Stone, Glass, Munn, Tugwell, & Doi, 2020). With the help of a Semantic Differential Scale (SDS), students and teachers can determine the test item difficulty. SDS is used on the basis that the scale provides a more detailed unit of measure, namely the scale range 1 to 7, where 1 and 7 are real numbers. This method can use the Focus Group Discussion (FGD) approach.

The approach is a way that can be done to solve a test item in a short but accurate time (Lafferty, 2004). Furthermore, the approach is efficient when using smaller groups (Lafferty, 2004; Guest, Namey, & McKenna, 2017). Research results used consensus theory by involving experts concluding that the characteristics of the test items are not different from computer program analysis (Kozierkiewicz-Hetmańska & Poniatowski, 2014).

Therefore, determining the test item's difficulty requires other ways according to the needs of teachers at school. Nevertheless, the difficulty is expected to be equivalent to the output of the program. This method can be achieved by comparing the program's outputs with the FGD approach, teacher group, student group, and teacher-student group with the help of the scale.

Literature Review

Test Item Difficulty

Characteristics of the test item difficulty refer to the continuous line ability of the examinee. Then, the difficulty is one of the characteristics of the right test item. There are several other characteristics of test items, such as discrimination power, distractor efficiency, validity, and test reliability, including scoring (Applegate, Sutherland, Becker, & Luo, 2019). The test item difficulty is the proportion of examinees' answers correctly compared with the examinee number. The proportion of test items is easy and hard. It can be refined extreme to be very hard, hard, medium, easy, and very easy. The difficulty is natural, although the substantial touch is difficult to practice in theory. In practice, it is often done for the provision of the test when the test is using a group approach. So, the assembly test requires more art than science.

Overview of TIMSS

Trends International Mathematics and Science Study (TIMSS) is an international-level study to see the trend of math and science abilities. TIMSS has two forms of test items, namely multiple-choice and construct response. Multiple choice test items are with four answer choices for grade IV elementary school students and five answer choices for grade VIII middle school students. For the test construct response items, the construct response test item has two forms, namely a brief construct response and a multilevel construct response. The test items are based on the ability framework in mathematics, which consists of dimensions and domains.

In the 2015 TIMSS assessment frameworks, assessment is divided into two dimensions, i.e., the dimensions of content and cognitive dimensions. Dimensions of content for fourth-grade students consist of three domains as follows. First, the content number of 50 percent is derived from the topic of numbers, fractions, decimal numbers, patterns, and relationships. Second, the geometric shapes and measurements of 35 percent are derived from the topic lines and corners, two and three dimensions, location, and motion. Third, the data presentation of 15 percent is derived from reading, interpreting, organizing, and representing. All the domains are associated with the assessment of the cognitive dimension.

Focus Group Discussion

Focus Group Discussion (FGD) relates to a method of collecting data from a study. FGD's usefulness is to obtain data/information from various sources and relevant preferences in a group discussion. The main objective of obtaining data interaction of a discussion group examinee / respondent is to increase the depth of information exposing various aspects of a phenomenon of life. FGD has several characteristics as follows. First, the approach is a method of collecting data for qualitative research. The data produced from the exploration of social interaction occurs when informants conduct the discussion process. Second, the implementation of the approach activities is carried out objectively and externally. Third, the approach needs a facilitator/moderator trained and reliable to facilitate discussion so that the interactions that occur among examinee focus on item-solving (Hennink, 2014). Fourth, the approach uses semi-structured interviews with a group of individuals with a moderator leading the discussion with informal arrangements and aims to collect data or information on a particular issue. Fifth, the approach has the characteristics of the number of individuals who are quite varied for a group discussion. One discussion group can consist of 5 to 8 people (Krueger & Casey, 2014). There are other variations, but not too large, so as to be more effective and easy managed to achieve the goal (Guest et al., 2017).

Methodology

This research used a quantitative-comparative type approach by comparing the result of estimated test item difficulty using SDS approaches with the test item difficulty from the Program of Item and Test Analysis. Both approaches consisted of four groups: FGD, teacher group, student group, and teacher-student group. So that teachers did not dominate the activities in the FGD group and the teacher-student group, the researchers explained to the teachers about the assessment process in the classroom, especially the assessment as learning and assessment for learning.

Participant

The study population was public elementary schools in Soppeng Regency, South Sulawesi, Indonesia. Sampling was done randomly, taking one village in Soppeng Regency. From the randomization results, Village Jennae was chosen. All elementary schools in the villages were included in the study. The sample size was 35 people composed of 14 teachers and 21 students.

Data Collection

The test item TIMSS was downloaded from the page https://timssandpirls.bc.edu/. The study only took a test item of multiple-choice type on mathematic, with as many as 40 items on the content and cognitive dimensions. The item test was validated in terms of content, language, and construction. The examinees' answers to the test item and SDS sheets were collected and processed in each group.

Semantic Differential Scale

SDS is a contrasting word instrument. The word resistance uses the smallest scale unit of 1, and the largest is 7. The numbers from 1 to 7 are real numbers. SDS is widely used to determine a certain point based on one's preferences for something. Hys and Hawrysz (2014) proposed the use of SDS to assess the advantages and disadvantages of QMS for certificate accreditation in Poland. Furthermore, Olaogun, Adedoyin, Ikem, and Anifaloba (2009) suggested that SDS was reliable and valid for assessing the symptom status of patients, namely, patients with low back pain. The SDS scale has seven units of assessment aid points. The SDS scale range is [1,7]. The values of the interval are the real number in the form of three decimal places. A score of 7,000 indicates that the test item has very great difficulty. Conversely, a score of 1,000 means that the test question has a very low problem difficulty. The teacher and students work on the test questions for 3 minutes and then estimate the test item difficulty with SDS for 1 minute.

Data Analysis

The estimated test item difficulty is divided into four, namely, estimated difficulty with FGD, estimated difficulty with the teacher, estimated difficulty with a student, and estimated difficulty with teacher-student, where the symbols are 1, 2, 3, and 4. In the same way, the program output has symbols 5, 6, 7, and 8.

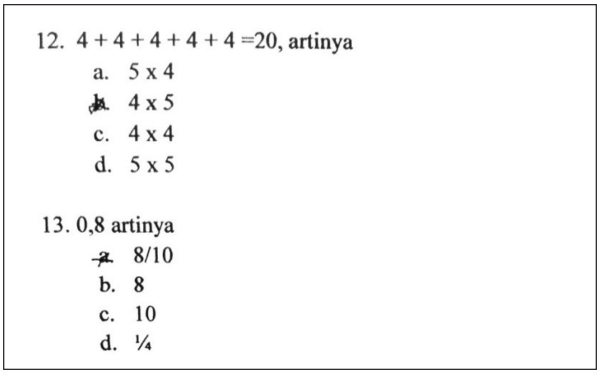
Program of Item and Test Analysis uses version 3.6. This version is not compatible with Windows 10, so it requires an additional program that is DOSOSBox0.74-win32-installer.exe. The program processes the response data of examinees for each group. Data that has been responded to by examinees is then entered into Excel. The file Excel is transferred to Notepad for each group (Guyer & Thompson, 2006).

The data analysis of research used nonparametric statistics, i.e., Kruskal-Wallis, for more than two groups. Furthermore, the two groups test can use the Mann Whitney test. The significance test uses a significance level of .01. Both tests are the conservative test for the small sample of the test, and the data group is not normally distributed (Salkind, 2013; Wallace, 2004). Software used for data analysis is, namely, SPSS Version 16 (SPSS Inc., Chicago, IL, USA).

Results

Description of Teacher and Student Activity

Figure 1 shows the work of one of the students from two test items. The test items were as follows, "4 + 4 + 4 + 4 + 4 = 20, the meaning is" and "0.8, the meaning is". Test item number 12 contained the concept of multiplication, which is 4 + 4 + 4 + 4 + 4 = 20. The choices were 5 x 4; 4 x 5; 4 x 4; or 5 x 5. The student chose 4 x 5 so was given a score of zero for a wrong answer. Test item number 13 contained the concept of transforming decimal to fraction. The test item required the examinee to change 0.8 to a fraction. The choices were 8/10, 10/8, 4/5, and ½. The student chose 8/10 so was given a score of one for a correct answer. The information has a quantitative value. The teacher realizes the difficulty and gives feedback about students' abilities regarding the characteristics of the test items.



12. 4 + 4 + 4 + 4 + 4 =20, artinya

    a. 5 x 4

    [✓ b.] 4 x 5

    c. 4 x 4

    d. 5 x 5

13. 0,8 artinya

    [✓ a.] 8/10

    b. 8

    c. 10

    d. ¼

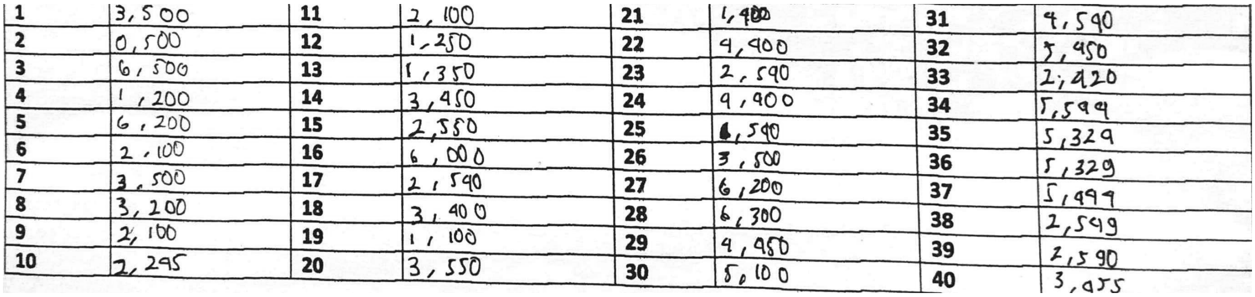
Figure 1 Student answers to test items

Figure 2 shows the results from estimates of test item difficulty of the FGD groups. The description of the test item estimation at SDS shows that the test item number 12 has a test item difficulty of 1.25, while test item number 13 has a test item difficulty of 1.35. So, both test items are in the easy category in the scale range [0,7] on SDS. Test item number 12 is related to the concept of multiplication process material. The student lacks understanding of the idea of the multiplication process so chose option b, not a.

Test item number 13 is more difficult for students to do than test item 12, but the answer was correct. The test item is related to the subject matter about the transformation of decimal form into fraction form. The value 0.8 as a decimal form is then transferred to the 8/10 fraction. The results of the second test item provide information for the teacher to give feedback on the previous material. The teacher can predict the abilities of these students to the material in the form of feedback. For the test items, which according to the students were easy, they answered incorrectly. On the other hand, for the test items, which according to the students were difficult, they answered correctly.

Comparison of All Groups

The distribution of data of each group is not the same, so it is more suitable to use the difference in the average rating than the median. The mean rank of the eight groups is different. However, these differences cannot provide meaningful, significant, or insignificant information, which requires another test. For the test of differences between more than two groups, the Kruskal-Wallis test can be used (Dalgaard, 2008; Vargha, Delaney, & Vargha, 1998).



1	3,500	11	2,100	21	1,400	31	4,590
2	0,500	12	1,250	22	4,400	32	5,450
3	6,500	13	1,350	23	2,590	33	2;420
4	1,200	14	3,450	24	4,400	34	5,599
5	6,200	15	2,550	25	1,590	35	5,329
6	2,100	16	6,000	26	3,500	36	5,329
7	3,500	17	2,540	27	6,200	37	5,499
8	3,200	18	3,400	28	6,300	38	2,549
9	2,100	19	1,100	29	4,450	39	2,590
10	2,295	20	3,550	30	5,100	40	3,455

Figure 2 The estimated value of the test item difficulty

Test of Kruskal-Wallis

The eight group difference tests showed that asymptotic significance is smaller than .01, meaning there are one or more different groups. So, there are groups that are the same, and some that are not the same. The results of the eight group difference test get the value of asymptotic significance smaller than .01. The assumption is accepted that there is a different group from the eight groups that exist, so it needs to be tested further for the differences between the two groups. Therefore, further testing is needed using the two-group comparison test. Several test options can be used for the test. However, specifically, for small samples and not the same sample size, the Mann Whitney test can be used (de Winter & Dodou, 2010). The eight group difference tests showed that asymptotic significance is smaller than .01, meaning there are one or more different groups. So, there are groups that are the same, and some that are not the same. The results of the eight group difference test get the value of asymptotic significance smaller than .01. The assumption is accepted; there is a different group from the eight groups that exist. It needs to be tested further for the differences between the two groups. Therefore, further testing is needed using the two-group comparison test. Several test options can be used for the test. However, specifically, for small samples and not the same sample size, the Mann Whitney test can be used (Nachar, 2008). However, this research only used groups using the scale for assessment test item difficulty. Next, further analysis is required for comparing with the program output according to respective response data.

Comparison of Four Group Pairs

The study compares the two groups' tests. This comparison uses the assumption that data from the two groups come from the same source. For this reason, the statistical test only tests four pairs of groups.

Group 1 and 5

The difference in the estimated difficulty of the test items from the test examinees from the FGD group uses the scale and program output using the test examinee's response data from the group.

As shown in Table 1, the asymptotic significance value is smaller than .01. So, the presumption is that there is a difference between groups 1 and 5. There is a difference between the estimated results using the scale approach and the program in the group. Thus, the test item difficulty from the estimated FGD using the scale is different test item difficulty from the output of the program, where the data comes from the response of the FGD.

Group 2 and 6

The difference in the estimated difficulty of the test items from the test takers from the teacher group uses the scale, and the program output uses the test examinee response data from the teacher group.

As shown in Table 2, the asymptotic significance value is smaller than .01. So, there is a difference between groups 2 and 6. That is a difference between the estimated results using the scale approach and the program in the teacher group. Thus, the test item difficulty from the estimated teacher group using the scale is different test item difficulty from the output of the program, where the data comes from the response of the teacher group.

Group 3 and 7

The difference in the estimated difficulty of the test items from the examinees from the student group uses the scale, and the program output uses the test examinee's response data from students.

Table 1 Differences; group 1 and 5

Test Statistics	
	Difficulty
Mann-Whitney U	334.000
Wilcoxon W	1154.000
Z	-4.484
Asymp. Sig. (2-tailed)	.000
a. Grouping Variable: Factor	

Note: p < .01.

Table 2 Differences; group 2 and 6

Test Statistics	
	Difficulty
Mann-Whitney U	234.000
Wilcoxon W	1054.000
Z	-5.466
Asymp. Sig. (2-tailed)	.000
a. Grouping Variable: Factor	

Note: p < .01.

As shown in Table 3, the asymptotic significance value of two groups of $p = .128$ is greater than .01. There is no difference between the estimated results using the scale approach and the program in groups of students. Thus, the test item difficulty from the estimated student group using the scale is different test item difficulty from the output of the program, where the data comes from the response of the teacher-student.

Group 4 and 8

The difference in the estimated difficulty of the test items from the test takers from the teacher-students group uses the scale and output of the program using test examinee response data from teacher-students. As shown in Table 4, the asymptotic significance value of two groups of $p = .03$ is greater than .01. So, there is no difference between the two groups. Thus, the test item difficulty from the estimated teacher-student group using the scale is different test item difficulty from the output of the program, where the data comes from the response of the teacher-student.

Discussion

The results of the study on the difficulty estimation of the test items show the following points. First, the examinee from the FGD does not have a similarity in difficulty with the output of the program. These findings provide information that the FGD is not acceptable to use to determine the difficulty of the test item. The results of the study also did not support the results of previous studies that the estimated characteristics of test items using FGD showed the similarity of the results of the application program output (Kozierkiewicz-Hetmańska & Poniatowski, 2014). Some of the obstacles influenced the results of such studies as follows. The teachers and students have never had an open and direct dialogue to solve problems together in FGDs. Besides, the study uses seven FGD groups, where each group has five members consisting of three students and two teachers from the same school. Therefore the small sample size is seven, too small to be analyzed on the program. Therefore, the output of the program is unstable. The result is the same as the teacher group.

Table 3 Differences; groups 3 and 7

Test Statistics	
	Difficulty
Mann-Whitney U	642.000
Wilcoxon W	1462.000
Z	-1.521
Asymp. Sig. (2-tailed)	.128
a. Grouping Variable: Factor	

Note: $p < .01$.

Table 4 Differences; groups 4 and 8

Test Statistics	
	Difficulty
Mann-Whitney U	575.000
Wilcoxon W	1395.000
Z	-2.165
Asymp. Sig. (2-tailed)	.030
a. Grouping Variable: Factor	

Note: $p < .01$.

Second, examinee from the student group shows that the difficulty from the estimated students using the scale is the same as the difficulty from the output of the program. Therefore, students can assess the difficulty degree of the test item. The number of students involved in the study was as many as 21 people, where the size is greater, so the representation of the assessment is broader and stable, although the estimation is less good when compared to the teacher's estimate.

Third, examinee from the teacher-student group showed that the difficulty of the test items from the estimated teacher-students using SDS is the same as the difficulty from the output of the program. Therefore, a combination of teacher and student can be an estimator from the difficulty of test items. The number of teachers and students involved in the study was as many as 35 people, where the sample size is larger than the group of teachers and students, so the estimation is broader and more stable. Based on the results of the four assessment groups, the difficulty of the test items using the scale approach shows that the estimator groups that are following the output of the program are student or teacher-student. The theoretical impact of the results of the study is that estimators of the difficulty of the test item can use groups of teacher-student. The combined student and teacher is closer to the spirit of assessment for learning (Clark, 2012).

For the practice in school, two choices can be used by teachers to assess the difficulty of test items, namely, groups of students, or joint teacher -student. However, the combined teacher and student group is a more acceptable choice because of the larger sample size of 35, meaning the output of the program will be more stable. Thus, the development of test items at the school level, especially related to the difficulty of the test items, can already be found by combining teachers and students with the help of the scale.

Likewise, the development of Computerized Adaptive Testing (CAT) at the school level can be more easily and openly applied because there are other alternatives in how to assess the difficulty of test items without using a computer application. Although the study still uses the CTT approach, the CTT concept is less suitable for developing CAT items banks where the development is more suitable for the IRT concept (Glas & Van der Linden, 2003; Petersen et al., 2016). Nevertheless, the results of this study were used to conduct studies on the comparison of the estimator of the difficulty according to CTT concept. However, if it is applied according to the concept of IRT, it only adds to the sample size of the estimator where the number of teachers and students as examinees is around 500 for the one-parameter logistic model or the Rasch model (Wright & Mok, 2004). Furthermore, the 2 or 3 parameter model requires about 1000 or more examinees (Fraley, Waller, & Brennan, 2000).

This study used the CTT approach to assess the difficulty of the items, which does not require a large sample size. In this study a small sample size of 21 students and 14 teachers from seven schools was used. The sample size was too risky to generalize the research results. The sample size of teacher-student group of 35 test participants could fulfil the element of sample size adequacy. Furthermore, the theoretical assumption of comparison between groups was not studied theoretically in this study but only based on a reasonable belief that the comparison was due to the same data source. This study only compared the results of the estimation of the test item difficulty, and the results of the program output did not involve more detailed research of the test item material and test item construction. However, the results of the study found that the teacher-student group can be a reference for schools and teachers to estimate the difficulty of the items. The estimate is another alternative for teachers and schools, besides the program.

Conclusion and Recommendation

Groups of students or teachers-students can be an estimator of the difficulty of test items at school by SDS approach. The combination of teacher and student is best used as an estimator because it can support the assessment for learning. Based on this, teachers and students in schools can use the SDS or program to assess the test item difficulty. The FGD as an estimator for the test item difficulty is different from the output of the program so the FGD is not a good estimator, although, as a theory, FGD is an approach to effectively find a solution. Therefore, researchers can use this approach by increasing the number of groups to about 30 units. Besides, the difficulty cannot be entered into the CAT test item bank because of the research using the CTT approach. Therefore, other researchers can use the method of assessing the difficulty of test items with the IRT approach to be applied to the CAT test item bank.

Conflict of Interest

There is no conflict of interest.

Acknowledgements

The research was funded by the 2019 Budget Internal Grant from the University of Muhammadiyah Makassar through the Research and Community Service Institute. Therefore, the team would like to thank the Rector of the University of Muhammadiyah Makassar.

References

Applegate, G. M., Sutherland, K. A., Becker, K. A., & Luo, X. (2019). The effect of option homogeneity in multiple-choice items. *Applied Psychological Measurement*, *43*(2), 113–124. doi: 10.1177/0146621618770803

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research, 2*, 189–193.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249. doi: 10.1007/s10648-011-9191-6

Dalgaard, P. (2008). *Analysis of variance and the Kruskal–Wallis test*. doi: 10.1007/978-0-387-79054-1_7

de Winter, J. C. F., & Dodou, D. (2010). Five-point likert items: T test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research and Evaluation, 15*, 1–12.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*(2), 350–365. doi: 10.1037/0022-3514.78.2.350

Glas, C. A. W., & Van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27(4)*, 247–261. doi: 10.1177/ 0146621603027004001

Guest, G., Namey, E., & McKenna, K. (2017). How many focus groups are enough? Building an evidence base for nonprobability sample sizes. *Field Methods*, *29*(1), 3–22. doi: 10.1177/1525822X16639015

Guyer, R., & Thompson, N. A. (2006). *User's manual for the ITEMAN ™ conventional item analysis program*. Retrieved from http://smp.ypk.or.id/mediaguru/upload/ download/ITEMAN Manual.pdf

Hys, K., & Hawrysz, L. (2014). Semantic differential as an assessment tool of (dis) advantages of QMS in the light of accredited certification in Poland. *Chinese Business Review*, *13*(1), 42–52.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. doi: 10.1177/0146621616664046

Juškaite, L. (2018). Comparison of the national diagnostic paper-based and online tests in natural science. *Proceedings of the International Scientific Conference*. *25–26*, 293–303. doi: 10.17770/sie2018vol1.3438

Kozierkiewicz-Hetmańska, A., & Poniatowski, R. (2014). An item bank calibration method for a computer adaptive test. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8397,* 375–383. doi: 10.1007/978-3-319-05476-6_38

Lafferty, M. I. (2004). Focus group interviews as a data collecting strategy. *Journal of Advanced Nursing*, *48*(2), 187–194.

Martinková, P., Štěpánek, L., Drabinová, A., Houdek, J., Vejražka, M., & Štuka, Č. (2017). Semi-real-time analyses of item characteristics for medical school admission tests. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, *11*, 189–194. doi: 10.15439/2017F380

Hennink, M. M.. (2014). *Focus group discussions (Understanding qualitative research)*. Madison Avenue, New York, NY: Oxford University Press.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology, 4* (1), 13–20. doi: 10.20982/tqmp.04.1.p013

Olaogun, M. O. B., Adedoyin, R. A., Ikem, I. C., & Anifaloba, O. R. (2009). Reliability of rating low back pain with a visual analogue scale and a semantic differential scale. *Physiotherapy Theory and Practice, 20*(2), 135–142. doi: 10.1080/ 09593980490453048

Petersen, M. A., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., Hammerlid, E., … Groenvold, M. (2016). Development of an item bank for computerized adaptive test (CAT) measurement of pain. *Quality of Life Research*, *25*(1), 1–11. doi: 10.1007/s11136-015-1069-5

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between Classical Test Theory and Item Response Theory. *Educational and Psychological Measurement*, *76*(2), 325–338. doi: 10.1177/0013164415576958

Krueger, R. A., & Casey, M. A. (2014). *Casey-Focus groups: A Practical Guide for Applied Research* (5th ed.). Los Angeles, CA: SAGE Publications Ltd.

Salkind, N. (2013). Kruskal-Wallis One-Way Analysis of Variance. In *Encyclopedia of Measurement and Statistics*. doi:10.4135/9781412952644.n245

Hys K., & Hawrysz L., (2014). Semantic differential as an assessment tool of (Dis) advantages of QMS in the light of accredited certification in Poland. *Chinese Business Review*, *13*(1), 42–52. doi: 10.17265/1537-1506/2014.01.005

Stone, J. C., Glass, K., Munn, Z., Tugwell, P., & Doi, S. A. R. (2020). Comparison of bias adjustment methods in meta-analysis suggests that quality effects modeling may have less limitations than other approaches. *Journal of Clinical Epidemiology*, *117*, 36–45. doi: 10.1016/j.jclinepi.2019.09.010

Vargha, A., Delaney, H. D., & Vargha, A. (1998). The Kruskal-Wallis Test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*. doi: 10.2307/1165320

Wallace, D. P. (2004). The Mann-Whitney Test. *Journal of the American Society for Information Science and Technology*. doi: 10.1002/asi.10347

Wright, B., & Mok, M. M. C. (2004). An overview of the family of Rasch Measurement Models. *Introduction to Rasch Measurement*, 1–24. Retrieved from http://www.statistica.unimib.it/utenti/lovaglio/overview rasch.pdf