Evaluating math test of students through different item functioning: Who’s taken benefit and harmed?

S Solihatun1, I B Rangka1,*, J Siregar1, C Marisa1, Z Niswati1, S Sutrisno1, A Irawan1, D Napitupulu1, L Tike3, A Tigor3, R Rahim4, H Nurdiyanto5, and I Ifdil6

1 Universitas Indraprasta PGRI, Jakarta, Indonesia
2 Indonesian Institute of Sciences, Banten, Indonesia
3 Universitas Halu Oleo, Southeast Sulawesi, Indonesia
4 Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia
5 STMIK Dharma Wacana, Lampung, Indonesia
6 Universitas Negeri Padang, West Sumatera, Indonesia

*itsarbolo.rangka@unindra.ac.id

Abstract. Our study intended to evaluate material on math test in elementary school. This study involved 63 elementary school students in Jakarta. We used a fraction number and geometry test as a parameter to know how strongly the substance of the test items to testing the students ability in performing the math test. The different item functioning (DIF) aspect in Rasch Model was applied to identify the existence of the bias on measurement to the tests that have been performed. The results of this study prove that the quality of the tests is excellent. Unfortunately, the students consistency of answered the test are inadequate. Furthermore, there is 1 item exposed bias on gender. Female students are the group that have benefits and vice versa, male students are a group that is disadvantaged in working on item No. 17 in math test. Differences in ability between male and female students in working on each item on math test is also an interesting focus that we short discussed in this article. These findings suggest that the structure of questions addressed by teachers during math test is unsuitable with students’ abilities and disadvantages students because of poorly recognizable students capacities.

1. Introduction
The last few decades, the transformation and improvement of assessment learning and abilities for students in schools are increasing so rapidly and progressively compound [1, 2]. One provocative issue is the effort to identify whether the items in the applied test experience a bias towards a specific group or demography [3]. The concept of item bias is also based on differences in each group towards the level of item difficulty and the actual ability of students in their groups or subgroups. Well, each group of students may have differences in expressing answers to a test; but this may not mean a bias [4]. An item in a particular test is declared to be biased if the item results in a difference or an unusual score between the group’s performance and other groups in specific norms [5-7]; or more precisely, the concept of bias in a test can be examined from five aspects; the five meanings are as follows: (a) mean group differences, (b) differential predictive validity, (c) differential item functioning, (d) differing factor structures of tests, and (e) unequal consequences of test use for various groups [8]. Refractive items surely change the quality of the instrument or test that has been constructed [9].
A study linked to Different Item Functioning (DIF) as a way to recognize a biased item that has been presented a number [10-13], and the most familiar is detecting certain DIFs against gender or specific racial [14]; especially in the context of assessment in education. This seems or sounds extremely plain, but it is very crucial in the principle of measurement involving humans as objects of measurement [15, 16]. Using of DIF to identify injustice for test participants and increase the validity of measurements [17] to avoid measures that are not standardized. As we know unstandardized measures lack specific directions governing their use, administration, scoring, and interpretation [18]. In the context of assessment, students should not be harmed or benefit when a test is provided. This psychologically can affect students' images of understanding their destiny with the study they are currently standing; trigger stressing for students [19], and may refusing to take lessons in school. Furthermore, the practice of DIF also encourages teachers to draw students' abilities objectively.

Normally, the test of learning outcomes in elementary schools is presented in multiple choice questions (MCQ). Unfortunately, MCQ tests tend not to measure students' cognitive abilities that are further complex [20]. Meanwhile, mathematics studies emphasize mastery of students' ability to think complexly [21, 22]. As an example, in geometry learning students are required to be able to be taught in the cognitive domains - knowing, applying, and reasoning on the property to set up certain spaces and/or specific fields [23]. Another example, in mathematical test items presented in narrative form; students are most likely confronted with "word problems" on a "mathematics test", this can cause the test item biased [24]. This is absolutely essential because "text" in an item also affects the performance of an item in the test that is produced [25]. Instead of looking at an estimate of students 'mathematical abilities, on the contrary the teacher gets a different reality where students' abilities are not recognized properly because mathematics tests are constructed to be unbalanced and poorer, it is understood that the tests given have disadvantaged certain groups of students. It is realized that very few test or instrument developers pay attention to the DIF study [15]. In our witness, teachers in elementary schools are not adequately trained in developing objective tests, because they are dominant using the raw score to measure whether students succeed or neglect in learning at school. Meanwhile, the quality of the tests provided still needs to be questioned.

This study tries to evaluated the material of mathematical tests, specifically in the fields of geometry and fractions number in elementary schools. More pointedly, we investigate the quality of mathematical tests, including the consistency of students during the test, and the potential of items that are harmful or benefit to various student groups.

2. Method

This research involved 63 students on fifth grade in elementary school in South Jakarta; consists of 33 female and 30 male students. Students take semester exams for mathematics subjects on May, 2018. The basic competency material investigated is (1) recognizing basic fractions, (2) comparing basic fractions, (3) explaining problems related to basic fractions (4) identifying with simple flat builds of properties or elements, (5) identifying varied types and wide angle, (6) calculates the girth of a square and a rectangle, and (7) deals with an issue referred to circumference, area of square and rectangle.

The math test design to Multiple Choice Questions (MCQ) and consists of 35 items. Preliminary screening results and class teacher reports revealed 4 exam items were interrupted because there were no correct answers choice. Furthermore, particularly 31 test items were entered in the analysis component of this research. Technically, math tests are performed by paper-based formats. Data analyzed by WINSTEPS version 4.01 Rasch computer programs [26] to recognize DIF evidences. All student actions in this research are a credential and no credit received by students from the schools for his/her participation in this research. These research datasets can be accessed via https://osf.io/3vsg2/ Open Science Framework [27].
3. Results and discussion

3.1. Initial checking of math test quality and students consistency

The results of the data analysis revealed that the quality of the measurements provided by the mathematics test was superb ($\alpha = .94$). Unfortunately, the consistency of students in taking on math tests is extremely poor ($\alpha = .47$). This is again supported by the analysis of the interaction between mathematical tests with students who were not satisfying ($\alpha = .50$).

On the other hand, we discovered that item information units, notably in the analysis of raw variance explained by measures only reached 40.6%. Although theoretically this means that the math test is valid, sadly the person explained (8.4%) and explained items (32.1%) in the applied mathematical test didn’t show up a balanced distribution. This reflects the measurement of math tests that have not covered students' ability; there are still 59.4% of measurement aspects that have not been reached by constructed mathematical tests. This finding is an initial indication if there is a disparity between the formulation of test items produced and the actual abilities of students.

We conducted further investigations to examine the Outlier Sensitive Responses (Outfit) that occurred in students and items in the math test. This is extremely important given that students and items that outfit directly give a decrease in quality to the measures collected.

### Table 1. Item outfit on math test (Calculated by $I=31$, $N=63$).

| # Items | Outfit MNSQ | # Items | Outfit MNSQ |
|---------|-------------|---------|-------------|
| 10      | 3.05        | 2       | .48         |
| 19      | 1.78        | 11      | .47         |
| 29      | 1.69        |         |             |

Measure MNSQ in Logit.

### Table 2. Students outfit on math test (Calculated by $I=31$, $N=63$).

| # Students | Outfit MNSQ | # Students | Outfit MNSQ | # Students | Outfit MNSQ |
|------------|-------------|------------|-------------|------------|-------------|
| 01P        | 3.54        | 48L        | 1.90        | 60P        | .49         |
| 44P        | 3.38        | 37L        | 1.75        | 10P        | .44         |
| 06L        | 3.28        | 51P        | 1.74        | 26L        | .47         |
| 32P        | 3.20        | 08P        | .49         | 39L        | .38         |
| 59L        | 2.01        | 09P        | .49         | 33L        | .38         |
| 14L        | 1.96        | 28L        | .49         | 41P        | .35         |

Measure MNSQ in Logit. $L=Laki$-laki (Male), and $P=Perempuan$ (Female).

Satisfactory conditions for measure if the MNSQ Outfit quality is in the range 0.5 - 1.5 Logit. Based on Table 1 there are 5 Outfit items. Outfit items provides us information that items did not pair with the pattern of responses produced by students. Furthermore, based on Table 2 there are 18 outfit students. Students who are affected with an outfit are a group of students who experience problems in answering to an applied test. The problem referred to in this case is careless students, cheating, and / or lucky guess. Complying with this, and considering the reasons for measurement, items and students whose outfits cannot be evaluated to determine the potential for DIF.

3.2. Analysis of different item functioning: Did students really benefit or be harmed?

Did the students actually got benefit or are harmed in the implementation of the math test provided by the classroom teacher in elementary school? This is an interesting challenge to solve in this research. After we finish the initial checking, the items and students that pair the ideal measurement model are obtained. From this point we estimate; do items experience bias towards specific groups of students. In this research we included gender (Male and Female) as a demographic variable which will be reviewed through DIF. To determine whether the items in a test have a bias in a specific respondent category, we measure (1) the probability value of DIF, and (2) Contrast DIF which aims to determine whether or not
the difference is meaningful (effect size) [15, 28]. In Rasch modeling, an item is categorized as having a bias if it has a probability value smaller than 0.05 (prob. < 0.05).

### Table 3. DIF value on math test items (Calculated by I=26, N=45).

| Item# | Prob. | Effect Size based on DIF Contrast | Item# | Prob. | Effect Size based on DIF Contrast |
|-------|-------|----------------------------------|-------|-------|----------------------------------|
|       |       | Male | Female |       |       | Male | Female |
| 3     | .3801 | .55  | -.55   | 21    | .3999 | -.20 | .20   |
| 6     | .5204 | .61  | -.61   | 22    | .6132 | .41  | -.41  |
| 7     | .5879 | .54  | -.54   | 23    | .5020 | .61  | -.61  |
| 8     | .7222 | .17  | -.17   | 24    | .3955 | -.53 | .53   |
| 9     | .8972 | .09  | -.09   | 25    | .2517 | 2.30 | 2.30  |
| 12    | .4257 | .30  | -.30   | 26    | .1022 | .99  | -.99  |
| 13    | .8917 | -.21 | .21    | 27    | .2846 | -.52 | .52   |
| 14    | .3875 | -.55 | -.55   | 28    | .2402 | 1.45 | 1.45  |
| 15    | .6359 | .26  | -.26   | 29    | .2903 | -.98 | .98   |
| 16    | .5355 | .30  | -.30   | 30    | .8289 | .10  | -.10  |
| J7   | .0148 | -1.76 | 1.76  | 33    | .6978 | .19  | -.19  |
| 18    | .1987 | -.22 | 2.22   | 34    | .5546 | -.18 | .18   |
| 20    | .6089 | -.21 | .21    | 35    | .3346 | .63  | -.63  |

Based on Table 3 indicate that Items No. 17 exposed to gender bias (prob. .0148 <0.5). If we refer to the effect size value of each gender group, it is seen that female students are the beneficiary group (DIF Contrast = +1.76), and vice versa, male students are the deprived group (DIF Contrast = -1.76). In the Rasch modeling symbol (+) indicates a student's high ability, and the symbol (-) indicates a student's low ability. In this section, we identify that male students find it easier to solve problems in item No. 17, but on the contrary female students are more difficult to solve problems in item No. 17.

Indonesian version:

17. Nama bangun datar yang memiliki 4 sudut siku-siku dan 4 sisi sama panjang adalah...
   a. Persegi Panjang  c. Segitiga
   b. Persegi  d. Lingkaran

(Translated)

17. The name of a flat build that has 4 right angles and 4 equal sides is ...
   a. Rectangle  c. Triangle
   b. Square  d. Circle

Furthermore, differences between the ability of male and female students in working on math tests were also undoubtedly established in this research. In general, male students have the ability of +0.029 logit higher than female students in working on the math test. However, if we seek at individual student performance; 40P students are identified to be students with the highest ability (+2.55 logit); on the other hand, 29P and 03L are students with the lowest ability among all students (-.39 Logit) (Please visit https://osf.io/3vsg2/). In the structure of the items used in the math test, item No. 7 is the item that is most difficult to answer by students, and vice versa, item No. 13 is the most easily answered item for all students.
4. Conclusions
This research has presented message that mathematics tests designed by classroom teachers in elementary schools are still questionable. Although we know that the test has an outstanding quality, but in reality, the level of seriousness and consistency of students in dealing with test items is still inadequate. In addition, there are 18 students who call for to be handled with individually because they have issues in solving the math tests offered. The presence of items that are biased in math tests has created its own consequences, especially female students benefit and vice versa male students are harmed. In the measurement context, in order to map students' abilities in learning, this requires to be reformed.

Acknowledgements
We thank to Mrs. Sujiah, S.Pd. for technical support and collected the data set in this research. We did not receive a grant from any institution in this research. We also declare no conflict of interest in this research.

References
[1] Tindal G 2002 Large-scale assessments for all students: Issues and options Large-scale assessment programs for all students: Validity, technical adequacy, and implementation 1-26.
[2] Snow R E and Lohman D F 1989 Educational measurement, 3rd ed.: American Council on Education) pp 263-331.
[3] Tatsuoka K K, Linn R L, Tatsuoka M M and Yamamoto K 1988 Differential Item Functioning Resulting From The Use of Different Solution Strategies Journal of Educational Measurement 25 301-19.
[4] Mellenbergh G J 1989 Item bias and item response theory International journal of educational research 13 127-43.
[5] Cleary T A 1968 Test bias: prediction of grades of negro and white students in integrated colleges Journal of Educational Measurement 5 115-24.
[6] Sandoval J 1979 The WISC-R and internal evidence of test bias with minority groups Journal of Consulting and Clinical Psychology 47 919.
[7] Flaugher R L 1978 The many definitions of test bias American Psychologist 33 671-9.
[8] Warne R T, Yoon M and Price C J 2014 Exploring the various interpretations of "test bias" Cultural Diversity and Ethnic Minority Psychology 20 570-82.
[9] Berry C M 2015 Differential Validity and Differential Prediction of Cognitive Ability Tests: Understanding Test Bias in the Employment Context Annual Review Organizational Psychology Organizational Behav. 2 435-63.
[10] Teresi J A, Holmes D, Ramírez M, Gurland B J and Lantigua R 2001 Performance of cognitive tests among different racial/ethnic and education groups: Findings of differential item functioning and possible item bias J. Ment. Health Aging 7 79-89.
[11] Su Y H, Chang J and Tsai H 2018 Using credible intervals to detect differential item functioning in IRT models. pp 297-304.
[12] Woods C M 2009 Empirical selection of anchors for tests of differential item functioning Appl. Psychol. Meas. 33 42-57.
[13] Steca P, Monzani D, Greco A, Chiesi F and Primi C 2015 Item Response Theory Analysis of the Life Orientation Test-Revised: Age and Gender Differential Item Functioning Analyses Assessment 22 341-50.
[14] Bond T G and Fox C M 2013 Applying the Rasch model: Fundamental measurement in the human sciences: Psychology Press).
[15] Boone W J, Yale M S and Staver J R 2014 Rasch analysis in the human sciences: Springer Netherlands).
Engelhard Jr G 2013 *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*: Routledge.

Shealy R and Stout W 1993 A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF *Psychometrika* **58** 159-94.

Gillespie D F and Perron B E 2015 *Key concepts in measurement*: Pocket Guides to Social Work R).

Alizamar A, Ifdil I, Fadli R P, Erwinda L, Churnia E, Bariyyah K, Refnadi R and Rangka I B 2018 The Effectiveness of Hypnotherapy in Reducing Stress Levels *Addictive Disorders & Their Treatment* Inpress.

Frederiksen N 1984 The real test bias: Influences of testing on teaching and learning *American psychologist* **39** 193.

Cargnelutti E, Tomasetto C and Passolunghi M C 2017 The interplay between affective and cognitive factors in shaping early proficiency in mathematics *Trends Neurosci. Educ.* **8-9** 28-36.

Moustafa A A, Tindle R, Ansari Z, Doyle M J, Hewedi D H and Eissa A 2017 Mathematics, anxiety, and the brain *Rev. Neurosci.* **28** 417-29.

Ubuz B and Aydin U 2018 Geometry knowledge test about triangles: evidence on validity and reliability *ZDM - Mathematics Education* **50** 659-73.

Shealy R and Stout W 1991 An item response theory model for test bias. Illinois University at Urbana dept of statistics).

Johnston P 1984 Prior knowledge and reading comprehension test bias *Reading Research Quarterly* **219-39**.

Linacre J M 2018 *Winsteps® Rasch measurement computer program* (Version 4.1.0). (Beaverton, Oregon: Winsteps.com).

Rangka I B 2018 Datasets: Evaluating test materials on fraction number and geometry for students through different item functioning. (New York: Open Science Framework).

Boone W J 2016 Rasch analysis for instrument development: Why, when, and how? *CBE Life Sci. Educ.* **15**.