Analysis of Multi Rater with Facets on Instruments HOTS of Solution Chemistry based on Tawheed

A Darmana¹, A Sutiani², H A Nasution², A Sylvia N.S³, N Aminah¹, T Utami³

¹,² Chemistry Department, Faculty of Mathematics and Natural Science, Medan State University, North Sumatra, Indonesia
³ Students of Chemistry Department, Faculty of Mathematics and Natural Science, Medan State University, North Sumatra, Indonesia

*Corresponding author email: ayidarmana2013@gmail.com

Abstract. This research is part of the development research in solution chemistry-HOTS-Tawheed. The purpose of this research is to develop of instrument test Higher Order Thinking Skills (HOTS) based on Tawheed on solution chemistry. The stages of the research began by designing 40 multiple choice item with five options. The instrument assessed by five science education experts. Expert judgement analysis of the test item was carried out using the Rasch measurement model approach, facets software. The Multi-rater validation test results showed: (1) The five expert panels considered quality items 23 out of 40 (57.5%); (2) according to the five validators, the instrument items that are most difficult to fulfill from the questions are statements B4, K4, and K8 (logit>+1), while K10 (logit<-4) is the statement item that is easiest to fulfill from the questions; (3) aspect of statistical fit criteria, it can be seen that Val.A and Val.F experts are the most consistent (based on MNSQ and ZSTD); (4) the reliability of the items showed a good level (0.97); and (5) the similarity of assessment by the five experts (exact agreement) which reached 93.0% which shows that there was not much difference between the five experts in assessing all the items.

1. Introduction
The Indonesian state has the goal of national education not only to achieve high intellectual aspects but also to include social and spiritual attitudes. This objective can be understood as a mandate of the law that all subjects in the school system in Indonesia must contribute to the achievement of national education goals. The objectives of national education as stated in Law No. 20 of 2003 aim to develop the potential of students to become faithful and righteous human beings. To God Almighty, have a noble character, are healthy, knowledgeable, capable, creative, independent, and become democratic and responsible citizens.

On the other hand, the results of PISA show that Indonesia has never obtained satisfactory results. Since its participation in 2000 until the implementation of the last PISA 2018. Of the three fields contested, namely reading, mathematics and science always score below average. PISA is held once in 3 years. PISA as an international competition event is believed to be credible enough to be used as an international standardization of education in Indonesia. PISA's low achievement is a barometer of the low achievement of national education goals. The scores for reading, math and science skills in 2018 were 371, 379, and 396 respectively. While the average score were 487, 489, and 489 respectively.
What is more concerning is that there is a decline from the acquisition in 2015 which scored 397, 386, and 403 for reading, math, and science skills, respectively. There have been many responses from researchers as an effort to improve PISA achievement as well as to increase its contribution to national education goals. This is shown by the many good studies that focus on the HOTS instrument [1-5] as well as to social and spiritual attitudes [6-12]. Research on efforts to improve scientific literacy [13-14]. Likewise, studies on the importance of presenting Islamic values for subjects other than religion in higher education [15]. The researchers themselves have conducted several studies from 2014 to 2019 which focus on the integration of religion and science, including the development of the HOTS-tawheed instrument [7], the development of learning media for science teaching materials integrated with spiritual values.

This study aims to develop a HOTS instrument based on the tawheed value in chemical solutions. Furthermore, the HOTS instrument was tested multiraterally (Multi facet Rasch Measurement. MFRM) using the review instrument. Specifically, this study aims to reveal: (1) which test items are good test items on the HOTS instrument, (2) the most difficult / difficult study items to fulfill by the HOTS instrument test items (3) Which validator is the most consistent in assessing HOTS instrument (4) inter-rater reliability (5) similarity level of validator's assessment

Basically, multirater analysis is an assessment of a test instrument based on expert judgment. Multirater analysis makes it possible to reveal several variables / factors that contribute systematically (non-randomly) to measurement variability [16]. The MFRM model is an extension of the basic Rasch model which is able to explain the variability that occurs due to aspects of the assessor, opportunity, or condition, in addition to item difficulty and the ability of test takers [17]. The MFRM model can also be used to explore the interactions between certain aspects. that MFRM is not only able to monitor the effect of the rater but also to detect systematic errors[18].

2. Methodology
This research design is an adaptation of some stages of research and development. The research design includes preliminary studies, design and development. In the preliminary stage, an analysis of teaching materials, general chemistry learning tools, especially solution chemistry, determining indicators, and a study of the concept of solution chemistry to compile of HOTS-tawheed instruments. Develop instrument of review for assessing HOTS-Tawheed instrument by adapted from [19]. This review instrument consists of three aspects of assessment, namely the material aspect (7 items), the construction aspect (10 items), and the language aspect (8 items). The total aspect of the item of the review instrument was 25 items.

At the design stage, based on the predetermined indicators, an instrument consisting of 40 HOTS items based on tawheed on chemical solutions was compiled. At the development stage, a multirater test was carried out on the HOTS instrument based on tawheed using the HOTS instrument review instrument. Involving 5 validators (4 lecturers majoring in chemistry and 1 teacher man 2 model. 2 out of 4 lecturers majoring in chemistry are lecturers in value education courses). Furthermore, analyzed using the Rasch model objective measurement approach with the Facets software.

3. Results and Discussion
The research stage began by designing as many as 40 multiple choice questions (PG) on the chemical solution with 5 alternative answers. Furthermore, expert judgment by 5 expert validators. Analysis of the test items was carried out using the Rasch model's objective measurement approach with the Facets software for expert validation testing. The following will present the results of the expert validator's analysis of the test items.

3.1 Result of expert judgment (Validation of Multirater)
The multirater validation description is presented in the following figure 1.

|Measr| Statements Item |Measr| Test item|Valid |
|-----|------------------|-----|---------|------|

2
In Figure 1, a multirater validity image is presented. The image consists of 4 columns, namely the 1st to the 4th column. The 1st column is the measure column (logit transformation) which represents the measurement result with a price range of +2 (top) to -8 (bottom), this number is / is called the logit value. In the 3rd column (item analysis by experts), describes the distribution of items that are in the logit value range from the smallest approaching -2 logit (items Q26 and Q30) to the greatest value approaching +2 logit (question items Q1). The logit price of 0 is the minimum criterion for the quality of questions considered to be of good quality by the expert, so that if the value is positive (greater than zero logit) it indicates a question that is considered good by the expert panel, whereas if the logit value is negative (less than 0) it indicates questions that are not good according to experts.

3.2 Validation of multiple choice test item

From Figure 1 above, it provides information that the items considered to be less good by experts are starting from item number 14 (Q14), 16, 23, 24, 38, 39, 4, 7, 17, 32, 33, 22, 31, 40, 6, 26, and 30 (17 items, 42.5%). While the items deemed qualified by the expert panel are items number 10 (Q10), 11, 13, 15, 25, 28, 37, 12, 19, 27, 3, 35, 5, 9, 21, 29, 34, 18, 20, 2, 36, 8, and 1 (there are 23 valid items, 57.5%).

Figure 1. Multirater Validation Test Results with Facets Software
3.3 Distribution of the difficulty level of validation instrument items

Information on the level of difficulty of the validation instrument statement items is shown in the 2nd column (item column) of Figure 1 above. The validator’s response to the validation instrument statement item is the validator’s assessment of the validation instrument statement item (item column). The column shows the distribution of the material aspects (M), construction (K), and language (B). M1 means that the material aspects are reviewed by the validator for the 1st statement (there are 7 statements for the material aspects, meaning that there are M1 – M7). The same thing for K, K1 means that the construction aspects are reviewed by the validator for the 1st statement (there are 10 statements for the construction aspect, meaning there are K1 – K10). B1 means that the language aspect is reviewed by the validator for the 1st statement (there are 8 statements for the language aspect, meaning that there are B1 – B8).

The validation instrument statement item that is reviewed with the lowest / most negative / lowest negative logit price shows the validation instrument statement item that is most easily fulfilled by the items according to the validator, while for the positive logit price, the more positive / bigger / upper it shows the statement item which is difficult to find on the items assessed by the expert panel. For item assessment items that have the same logit price means having the same level of ease / difficulty. From the column item Figure 1 above, information is obtained that B4, K4, and K8 (the logit price is the greatest / greater than +1) are the most difficult / difficult statement items to fulfill from the items assessed according to the five validators. Meanwhile, K10 (the lowest logit price / less than -4) is the easiest statement item to be fulfilled from the items according to the five validators. K10 is a statement about "answer choices in the form of numbers arranged according to the order of the size of the number", this shows that almost all the items whose answers are in the form of numbers fulfill this aspect of assessment. Whereas statements with a value of more than +1 logit, namely B4 (the formulation of each item uses a question word that demands an answer), K4 (the subject matter does not use a double negative statement), and K8 (there is one correct answer) shows that all the items according to the five expert panels have not fulfilled this aspect.

3.4 Level of validator ease

In the 4th column (validator) the distribution of the five validators is at negative logit values (-2 to -7). This provides information that the five validators are including the assessors who are not so strict in giving an assessment of the items given. The order from the evaluator with the most stringent way of evaluating to the least expensive rate is Val.A, Val.F, Val.J, Val.S, and Val.I.

Table 1. Expert Panel Validation Test Results with Facets Software

| Validator | Measure | Model S.E. | Infit MnSq ZStd | Outfit MnSq ZStd | Correlation PtMea |
|-----------|---------|-----------|----------------|-----------------|------------------|
| Val.A     | -2.45   | .12       | 1.14           | 1.8             | 1.35             | 1.7              | .48              |
| Val.F     | -4.23   | .19       | .82            | -1.2            | .56              | -1.0             | .47              |
| Val.J     | -4.71   | .21       | .96            | -2              | .48              | -9               | .42              |
| Val.S     | -5.42   | .25       | .64            | -2.3            | .18              | -1.3             | .43              |
| Val.I     | -7.26   | .47       | 1.04           | .2              | 8.15             | 2.1              | .12              |

Separation 5.62  Strata 7.82  Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 250.0  d.f.: 4  significance (probability): .00
Inter-Rater agreement opportunities: 10000
Exact agreements: 9298 = 93.0%  Expected: 9263.7 = 92.6%

Table 1 describes the quality of the assessment by a panel of experts sorted from the severity of the items. From the aspect of the statistical fit criteria [17], it can be seen that the Val.A and Val.F experts are the most consistent (the MNSQ Outfit and ZSTD Outfit values are both in the range of statistical suitability, namely MNSQ in the range 0.5 - 1.5 and ZSTD at range -2 to +2). In addition, the control limit for the quality of the assessment by panel experts is also determined based on the "infit" and
"outfit" values in the Rasch analysis, which are between 0.6 and 1.4 [20]. The worst expert was Val.I when the HOTS instrument assessment of tawheed based solution chemistry was used, while the most generous expert assessor was Val.A on all the instruments being assessed. From the aspect of inter-rater reliability, the score is very good (0.97), this shows that the experts give different scores [21]. The rater's tendency to influence the value of reliability [22]. The resulting data was in accordance with the measurement model which was shown to be significant from the Chi-square test value (p <0.01). What is interesting is the similarity in the assessment by the five experts (exact agreement) which reached 93.0% which shows that there is not much difference between the five experts in assessing all the items. Several studies in the literature also reveal that the rater's assessment tendency may be different and the rater's reliability is influenced by the behavior of the assessors such as leniency and severity [23-25].

4. Conclusion
The results of the analysis of the HOTS test instrument for tawheed based solution chemistry with MFRM showed interesting and more complete results that could provide more precise information about the quality of the questions. Multirater assessment which is intended to avoid bias can be executed more precisely through MFRM where the interaction of item-items, items and rater is calibrated on an equal-interval scale. The results of the analysis showed that (1) there were 17 out of 40 items deemed not good by the expert (42.5%), while 23 of the 40 items (57.5%) were considered good quality; (2) according to the five validators, the instrument items that were most difficult / difficult to fulfill from the questions were B4, K4, and K8 (the logit price was the biggest / greater than +1), while K10 (the logit price was the lowest / less than -4) is a statement item that is easiest to fulfill from the items; (3) From the aspect of the statistical fit criteria, it can be seen that the Val.A and Val.F experts are the most consistent (the MNSQ Outfit and ZSTD Outfit values are both in the fit item range); (4) the inter-rater reliability obtained a very good value (0.97) and the data was in accordance with the measurement model which was shown to be significant from the Chi-square test value (p <0.01); (5) the similarity of the assessment by the five experts (exact agreement) which reached 93.0% which shows that there was not much difference between the five experts in assessing all the items.

Reference
[1] Hanifah, N. (2019). Pengembangan Instrumen Penilaian Higher Order Thinking Skill (HOTS) di Sekolah Dasar. Current Research in Education: Conference Series Journal. 1(1),1-8.
[2] Anisah dan Lastutii,S.(2018, Desember). Pengembangan Bahan Ajar berbasis HOTS untuk Meningkatkan Kemampuan Pemecahan Masalah Matematis Mahasiswa. Jurnal Matematika Kreatif-Inovatif. 9(2), 191-97. [http://journal.unnes.ac.id/nju/index.php/kreano]
[3] Jonny, H.P., Rajagukguk, D. and Rajagukguk, J., 2020, January. Computational Modelling Based on Modellus to Improve Students’ Critical Thinking on Mechanical Energy. Journal of Physics: Conference Series (Vol. 1428, No. 1, p. 012042). IOP Publishing.
[4] Nisa, N. A. K., Widyastuti, R., & Hamid, A. (2018, July). Pengembangan Instrumen Assessment Higher Order Thinking Skill (HOTS) pada Lembar Kerja Peserta Didik Kelas VII SMP In Prosiding Seminar Nasional Matematikadan Pendidikan Matematika (Vol. 1, No. 2, pp. 543-556).
[5] Anas, L.H., Rajagukguk, J. and Bunawan, W., 2020, March. Video Technology Media based on Heat and Temperature to Improve of Learner Critical Thinking. Journal of Physics: Conference Series (Vol. 1485, No. 1, p. 012037). IOP Publishing.
[6] Assysyifa, D.S., Sopyan, A., Masturi. (2017). Pengembangan Bahan Ajar Ipsa Berbasis Komplementasi Ayat-Ayat Sains Quran pada Pokok Bahasan Sistem Tata Surya. Unnes Physics Education Journal.6 (1) .44-54.
[7] Darmana, A., Sutiani, A., Jasmidi (2020). Development of The Thermochemistry-HOTS-Tawheed Multiple Choice. Journal of Physics Conference Series.1462 :012051

[8] Faizah, N.S., Mubin, M. (2018). Pengaruh Modul Tematik Berbasis Islam Dan Sains Pada Tema Energi Dan Perubahannya Terhadap Hasil Belajar Siswa Mi Murni Sunan Drajat Lamongan. JPPiPA (Jurnal Penelitian Pendidikan IPA). 3 (2). 1-5 Introduction. Seventh edition, Boston: Pearson Education Inc

[9] Khoiri, A., Agus suryani, Q., dan Hartini, P. ( 2017). Penumbuhan Karakter Islami melalui Pembelajaran Fisika Berbasis Integrasi Sains-Islam. Jurnal Keguruan dan Ilmu Tarbiyah. 02 (1). 19-31

[10] Kurniasih, W.(2016). Implementasi Integrasi Agama Dan Sains Dalam Pembelajaran Matematika Di Sekolah Dasar Islam Terpadu Harapan Umat Brebes, Dialektika P. Matematika. 3(2) .20-29

[11] Pernadi, B.A.( 2018).Pengembangan Modul IPA Berbasis Integrasi Islam Dan Sains Untuk Meningkatkan Hasil Belajar Siswa Kelas VI Min 2 Mojokerto, Jurnal Pendidikan Islam. 1 (2) . 294-311

[12] Wastyanti, A.(2016). Pengembangan Bahan Ajar Integrasi Sains Dengan Agama Berbasis Multimedia Di Madrasah Ibtidaiyah Negeri Malang 2.Tesis, Pascasarjana Universitas Islam Negeri Maulana Malik Ibrahim Malang

[13] Sutiani, A., Zainuddin, Darmana, A., Panggabean, F.T.M. (2020) The Development of Teaching Material Based on Science Literacy in Thermochemical Topic. Journal of Physics Conference Series.1462 :012051

[14] Sadhu, S., &Laksono, E. W. (2018). Development and Validation of an Integrated Assessment for Measuring Critical Thinking and Chemical Literacy in Chemical Equilibrium.International Journal of Instruction, 11(3), 557-572.

[15] Purwanto, Y., Saepudin, E., Hidayatullah, D.S., Darmana, A., Yani, E.A. (2020)Islamic character education and nationalism insight of students in the era of industry 4.0.International Journal of Psychosocial Rehabilitation 24, 5471 -5485

[16] Eckes, T. (2015). Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluation Rater-Mediated Assessments. Frankfurt: Peter Lang Edition.

[17] Boone, W.J., Staver, J.R., dan Yale, M.S. (2014). Rasch Analysis in the Human Sciences. Dordrecht: Springer.

[18] Downing, S. M. (2005). Threats to the Validity of Clinical Teaching Assessments: What About Rater Error? Medical Education, 39,350-355.

[19] Sani, R.A. (2019). Cara Membuat Soal Hots. Tangerang: Tira Smart.

[20] Wright, B. D. & Linacre, J. M. (1994). Reasonable Mean-Square Fit Values. Rasch Measurement: Transactions of the Rasch Measurement SIG, 8(3), 370.

[21] Koçak, D. (2020). Investigation of Rater Tendencies and Reliability in Different Assessment Methods with Many Facet Rasch Model. International Electronic Journal of Elementary Education, 12(4), 349-358.

[22] Black, P. (1998). Testing: Friend or Foe? London: Falmer PressBond, T.G. dan Fox, C.M. (2015). Applying The Rasch Model, Fundamentals Measurement in the Human Sciences (3rd edition). New York: Routledge.

[23] Güler, N. (2014). Analysis of Open-ended Statistics Questions with Many Facet Rasch Model. Eurasian Journal of Educational Research, 55, 73-90.

[24] Brookhart, S.M. & Walsh, J.M., Zientarski, W.A. (2006). The Dynamics of Motivation and Effort for Classroom Assessment in Middle School Science and Social Studies.Applied Measurement in Education, 19(2), 151-184. http://dx.doi.org/10.1207/s15324818ame1902_5

[25] Mulqueen C., Baker D. &Dismukes, R.K. (2000). Using Multifacet Rasch Analysis to Examine the Effectiveness of Rater Training.Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans.