Eğitimde Kuram ve Uygulama (EKU)
Journal of Theory and Practice in Education
ISSN: 1304-9496

2022, 18(2), 27-42

Evaluations of Turkish Science Teacher Curriculum with Many-Facet Rasch Analysis

Ilgım Özergun[1], Fatih Doğan[2], Göksel Boran[3], Serdar Arcagök[4]

[1] *Department of Mathematics and Science Education, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, ilgim.ozergun@comu.edu.tr*
[2] *Department of Mathematics and Science Education, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, fatihdogan@comu.edu.tr*
[3] *Department of Computer & Instructional Technologies Education, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, gboran@comu.edu.tr*
[4] *Department Of Elemantary Educaıion, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, serdar_arcagok21@comu.edu.tr*

Corresponding Author: Ilgım Ozergun

Article Type: Research Article

To Cite This Article: Özergun, I., Doğan, F., Boran, G., & Arcagök, S. (2022). Evaluations of Turkish science teacher curriculum with many-facet rasch analysis. *Eğitimde Kuram ve Uygulama, 18*(2), 27-42. https://doi.org/10.17244/eku.1180825

Ethical Note: Research and publication ethics were followed. In this study, the data were collected before 2020, and voluntary participation of study group was observed during the data collection period.

Fen Bilgisi Öğretmenliği Programının Çok Yönlü Rasch Analizi ile Değerlendirilmesi

Ilgım Özergun[1], Fatih Doğan[2], Göksel Boran[3], Serdar Arcagök[4]

[1] *Matematik ve Fen Bilimleri Eğitimi Bölümü, Eğitim Fakültesi, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, Türkiye, ilgim.ozergun@comu.edu.tr*
[2] *Matematik ve Fen Bilimleri Eğitimi Bölümü, Eğitim Fakültesi, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, Türkiye, fatihdogan@comu.edu.tr*
[3] *Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, Eğitim Fakültesi, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, Türkiye, gboran@comu.edu.tr*
[4] *Temel Eğitim Bölümü, Eğitim Fakültesi, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, Türkiye, serdar_arcagok21@comu.edu.tr*

Sorumlu Yazar: Ilgım Özergun

Makale Türü: Araştırma Makalesi

Kaynak Gösterimi: Özergun, I., Doğan, F., Boran, G., & Arcagök, S. (2022). Evaluations of Turkish science teacher curriculum with many-facet rasch analysis. *Eğitimde Kuram ve Uygulama, 18*(2), 27-42. https://doi.org/10.17244/eku.1180825

Etik Not: Araştırma ve yayın etiğine uyulmuştur. Bu çalışmada veriler 2020 yılı öncesi toplanmış olup, veri toplama sürecinde katılımcıların gönüllü katılımı gözetilmiştir.

Eğitimde Kuram ve Uygulama (EKU)
Journal of Theory and Practice in Education
ISSN: 1304-9496

2022, 18(2), 27-42

Evaluations of Turkish Science Teacher Curriculum with Many-Facet Rasch Analysis

Ilgım Özergun[1], Fatih Doğan[2], Göksel Boran[3], Serdar Arcagök[4]

[1] Department of Mathematics and Science Education, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, ilgim.ozergun@comu.edu.tr, ORCID: 0000-0002-2277-6016
[2] Department of Mathematics and Science Education, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, fatihdogan@comu.edu.tr, ORCID: 0000-0002-3088-886X
[3] Department of Computer & Instructional Technologies Education, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, gboran@comu.edu.tr, ORCID: 0000-0003-3060-3876
[4] Department Of Elemantary Educaıon, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, serdar_arcagok21 @comu.edu.tr, ORCID: 0000-0002-4937-3268

Abstract

Scientific and technological developments cause changes in educational programs and curriculums. Especially science education should meet criteria of today's needs and expectations. Changing only science curriculum in K-12 is not enough. Science teacher curriculum should also change since teachers are responsible to teach subjects. By 2018, all teacher curriculum, including science teacher education, changed due to recent improvements in science, technology and education. This study investigated science teacher educators' evaluations of Turkish science teacher curriculum with Many Facet Rasch Analysis. The program is evaluated according to the four dimensions of curriculum which are 1) aims, aims objectives, 2) subject matter, 3) learning experiences, and 4) evaluating approaches. These analyses including general evaluations about the program, academicians' generosity, and ungenerosity behavior during evaluating the program, and analysis of each criterion itself. Results of the analysis conformed psychometric and unidimensional properties of the criterion form. Therefore, it is supported with the literature that a Likert-type instrument can be developed and used to evaluate programs. Additionally, this study discussed academician's generosity and ungenerosity behavior while evaluating the program. Evaluating validity and reliability of each academicians' behavior is necessary. Results indicated that their bias, generosity, or ungenerosity behaviors did not affect the criterion forms' statistical confidence.

Article Info

Keywords: Science teacher education, science teaching, program evaluation, many-facet Rasch analysis

Article History:
Received: 27 September 2022
Revised: 10 October 2022
Accepted: 10 October 2022

Article Type: Research Article

Fen Bilgisi Öğretmenliği Programının Çok Yönlü Rasch Analizi ile Değerlendirilmesi

Öz

Bilimsel ve teknolojik gelişmeler eğitim ve öğretim programlarında değişikliklere neden olmaktadır. Özellikle fen eğitimi, günümüzün ihtiyaç ve beklentilerini karşılamalıdır. Bunları karşılamak için sadece ortaokul fen müfredatını değiştirmek yeterli değildir. Öğretmenler, konuları öğretmekle sorumlu oldukları için fen bilgisi öğretmenliği eğitim programları da günümüzün ihtiyaçlarına göre değişmelidir. 2018 yılında, fen bilgisi öğretmenliği programı da dâhil olmak üzere tüm öğretmen eğitimi programları; bilim, teknoloji ve eğitimdeki son gelişmeleri programa dâhil etmek için değişti. Bu amaçla, bu çalışmada fen bilgisi öğretmenliği programı, fen bilgisi eğitimi programında çalışan akademisyenlerce değerlendirilmiş ve bu değerlendirmeler Çok Yüzeyli Rasch Analizi ile incelenmiştir. Program, 1) amaç ve hedefler, 2) konu, 3) öğrenme deneyimleri ve 4) ölçme ve değerlendirme olmak üzere dört program boyutuna göre değerlendirilir. Programla ilgili genel değerlendirmeler; akademisyenlerin programı değerlendirme sırasındaki cömertlik ve cimrilik davranışlarını göstermiş ve her bir kriterin kendi analizini ayrı ayrı göstermiştir. Analiz sonuçları, ölçüt formunun psikometrik ve tek boyutlu özelliklerine uymaktadır. Bu nedenle bu çalışmada geliştirilen Likert tipli ölçme aracının fen bilgisi öğretmenliği programının değerlendirilmesinde kullanılabileceği söylenebilir. Ayrıca, bu çalışmada akademisyenin programı değerlendirirken cömertlik ve cimrilik davranışları ele alınmıştır. Her akademisyenin davranışının geçerlilik ve güvenirliğinin ayrı ayrı değerlendirilmiştir. Sonuçlar, yanlılık, cömertlik veya cimrilik davranışlarının ölçüt formlarının istatistiksel güvenini etkilemediğini göstermiştir.

Makale Bilgisi

Anahtar Kelimeler: Fen bilgisi öğretmeni eğitimi, fen öğretimi, program değerlendirme, çok yönlü Rasch analizi

Makale Geçmişi:
Geliş: 27 Eylül 2022
Düzeltme: 10 Ekim 2022
Kabul: 10 Ekim 2022

Makale Türü: Araştırma Makalesi

İletişim/Contact: ilgim.ozergun@comu.edu.tr
DOI: https://doi.org/10.17244/eku.1180825

Introduction

Scientific and technological developments have one of the most important roles in shaping globalized world's needs. School curriculums are driving force behind societies' catching up expectations of 21st century (Bencze & Carter, 2011). For this reason, integrating science, technology, and society relationship into school curriculums are required. Teachers are responsible for teaching school curriculums. While teachers prepare students to future world, preservice teacher curriculum prepare future teachers who will raise next generations. Accordingly, there has been a great emphasis on teacher curriculum over years. For this reason, curriculums of teacher curriculum have been developing by taking societies and the world's needs into consideration (Bencze & Hondson, 1999). Developments in science and technology and expectations of society best fit science curriculums (Aikenhead, 1997). Science education begins with kindergarten years and continues through all K-12 years. It is expected science teachers to develop themselves and educate students for the need of today's globalized world (Ruggiero & Mong 2015).

Scientific, technological, and new trends in education have caused to major change for science teacher curriculum. In Türkiye, teacher curriculum are dependent on the Council of Higher Education (CoHE). Türkiye has given great importance to teacher curriculum in order to improve and develop them. By 2018, one of the biggest changes are applied to teacher curriculum by considering 21st century need, globalization of the world, and new trends in education (CoHE, 2018). In the new teacher curriculum CoHE provides names, terms, objectives, and contents for each course courses. In Türkiye, science teacher education is four-year undergraduate program which takes place in faculty of education. In the program there are compulsory and elective courses related to general sciences (physics, chemistry, biology), teacher pedagogy, and field (science teacher education). Since teacher pedagogy and science teacher education field courses are directly related to teaching approaches major changes happened at these courses. With these changes it is aimed to follow today's globalized, scientific and technological age.

Theoretical background

Teacher Curriculum

Teaching and learning are both main features of humankind. While subjects explored which are needed to teach and learn in the history, teaching has started to become a profession. After industrial revolution, scientific and technological developments have brought up teacher education. Previously, teachers were raised at teaching vocational high schools. By the time goes on, education faculties have been established to raise teachers (Okcabol, 2005). For example, in Türkiye, education faculties were established by 1982 and previous teacher education institutions incorporated to universities' faculty of education (CoHE, 2007). After that, teacher education in educational faculties have always been a changing and an improving area. There is no doubt that effective teacher education requires standards which meet needs of the age (Orhan, 2017). These standards are related to the general teaching competencies including knowledge about program, content, pedagogy, and discipline-based teaching. Similarly, most of the countries', including Türkiye's, teacher curriculum aim to provide skills related to humanities and social sciences, teaching as a profession and field of teaching itself (Popkewitz, 1994; Robinson & Latchem, 2003). Teachers who raise future generations who will be members of the society. Teacher education not only concentrates on subjects but also emphasis on today's and future's needs (Wei, 2020). Therefore, discipline specific teacher curriculum are needed. Universities' educational faculties accommodate different departments with programs. Mathematics and science education department offers science teacher curriculum.

Science Teacher Curriculum

Most of the countries' science teacher curriculum aims to improve preservice science teachers' content knowledge about science disciplines (such as physics, chemistry, and biology), pedagogical teaching strategies, and teaching methods of science through theoretical background and practical implementation during undergraduate years (Atkin, 1998). However, innovations in science and technology enforce science teaching to change and develop. Accordingly, Unal et al. (2004) investigated science education development progress in Türkiye. They argued that programs should be developed by considering previous programs insufficiencies, developments in science and technology, and needs of societies. Therefore, both science education in K-12 and science teacher curriculum are subject to change repeatedly. Moreover, science teacher curriculum should be compatible to K-12 science program. Therefore K-12 science and science teacher curriculum have to work together for successful science teacher education and science teaching.

CoHE (2018) claimed that educational faculties focus more on science content knowledge than science teaching strategies. This might cause to incorporation between K-12 science and science teacher curriculum. Cronin-Jones (1991) reported importance of teacher knowledge and belief on the implementation of the science teacher curriculum. One of the most significant knowledge is related how to teach specific subject to a particular age group.

This is both related to pedagogy of teaching and strategies of discipline specific teaching (Cronin-Jones, 1991). Pre-service teachers can learn and develop their teaching pedagogy and strategies through education faculties. These pedagogy and strategies can change over time due to the new century, developments in science and technology, and new generation. For this reason, Bawane and Spector (2009) argued that teacher curriculum should be change by also considering teachers' and future teachers' needs. Pre-service teachers' expectations and needs are important to improve teacher curriculum because they are future students. In Veal's (2004) study, pre-service teachers expect teacher curriculum to more content- and process-based programs which have authentic assessment techniques which are suitable with contemporary science education. Then pre-service teachers might feel closer to their students.

The Necessity of Improving Science Teacher Curriculum

Even though there is need for curriculum improvement, there was no change or developments in more than ten years. For this reason, it can be said that old programs were out-of-date. The factors that cause the program to be updated according to the results of the research and evaluation of the teaching undergraduate programs (Bawane & Spector, 2009). Teachers need modern, rich, globalized and up to date curriculum which includes lessons about content knowledge, teaching pedagogy, and discipline specific teaching and learning strategies. By these changes Türkiye's educational faculties, be prepared to globalized world by presenting modern program to teacher curriculum. Lessons' contents should include student centeredness, process and result oriented assessment techniques in educational faculties (MoNE, 2018). For this reason, CoHE of Türkiye has been updated to undergraduate level teacher curriculum by 2018-2019 academic year so that modernizing and adapting programs into today's world.

Teacher curriculum should be compatible with K-12 programs, too. Therefore, CoHE and MoNE have to work together in program and curriculum development. MoNE (2018) stated in the report that even though K-12 curriculum has changed over years, teacher curriculum' curriculum did not change over ten years. For this reason, teacher curriculum did not meet the age's criteria. Then CoHE changed all teacher curriculums by 2018. The aim of this change is to make compatible teacher curriculum with K-12 curriculum and needs. Science teacher curriculum is one of the updated teacher curriculum in 2018. These updates can be evaluated as dimensions of a program curriculum

Dimensions of a Program Curriculum

A curriculum is a designed set of course or content taught at a school or a university whereas a program is a set of structured activities of the curriculum. A curriculum is more comprehensive than a program. The dimensions of curriculum can be used to evaluate a program. (Olivia, 1997). Each curriculum has 4 dimensions; 1. aims, goals, and objectives, 2. subject matter, 3. learning experiences, and 4. evaluating approaches (Ornstein & Hunkins, 2009). The first dimension, *aims, goals and objectives* concentrate on expected statements of the observable action. While aim is the most general statement, goal indicates more specific outcomes and expectations, and objectives is the most specific observable action. The second dimension is *subject matter* which indicates contents to be taught. This dimension related to selection of activities, identification of topics and organizing experiences. Third dimension is *learning experiences* which concentrates on process of selecting and organizing of learning experience design. The last dimension is *evaluating approaches* relies on assessment and evaluation strategies. Evaluating approaches are divided as formative and summative assessment. For a curriculum or a program evaluation these four dimensions should be considered.

Many-Facet Rasch Analysis

The Rasch Analysis (1960) is a theory-based valid and reliable statistical probabilistic approach while developing, monitoring, or managing an instrument. This approach provides a probable illustration to researchers with regards to a criterion of the instrument or a participant of the study. Many-Facet Rasch Analysis (MFRA) categorizes ordinal and ratio scales to data which are beneficial to direct comparison for measurements. In other words, Many-Facet Rasch Analysis (MFRA) provides researchers with invariant scale to each criterion of the instrument so that latent trait remains the same. Therefore, MFRA is widely used when a comparison of criteria or bias of participants might affect the validity or reliability of the instrument. In other words, Rasch analysis used when researchers need to compare or contrast item and person reliability (Boone & Scantlebury, 2006).

Most of the participants of educational research are persons or documents. This nature of educational research might decrease reliability of the research or instrument itself. For this reason, Rasch Analysis is widely used in educational fields in last decades. Nature of MFRA allows researchers to analyze both large-scale and small-scale data. Many educational researchers around the World have used MFRA to evaluate large-scale assessments like PISA and TIMMS, instrument development and evaluation (Boone et al., 2011; Oon & Fan, 2017; Neuman et al., 2011), science education (Boone & Scanlebury, 2006; Jüttner et al., 2013; Bailes & Nandakumar, 2020). For example, You (2016) in the research developed a survey for science teaching practices. It is reported in the research that MFRA measures different aspects of content validity so that providing to construct valid and reliable forms. Similarly, Jüttner and colleagues (2013) suggested using MFRA while all respondent evaluating the same scale. This feature of MFRA enables survey development and future use of these surveys.

Boone et al. (2011) claimed that Rasch Analysis has a strong quantitative approach, however it should be used when a research problem needs qualitative analysis and approach. Since the problem statement of this study is appropriate for the nature of Rasch analysis, it is used. In this study two-facet, the Rasch Model design was used in order to analyze jury members' evaluations of criteria. Accordingly, both facet scores of criteria and jury members are calculated, independently. Baker (2001) suggested that before conducting Rasch Analysis, three assumptions should be provided which are (a) unidimensionality, (b) data-model fit, and (c) local independence.

a) Unidimensionality

Unidimensionality is a mode factor for assessing the purposeful psychological feature defined by Hambleton, Swaminathan and Rogers (1991). Unidimensionality is needed to compare the data is valid or not. For this reason, before interpreting the results, unidimensionality should be checked. Exploratory Factor Analysis (EFA) is used for unidimensionality of the criteria survey. EFA is a kind of unidimensionality analysis technique while finding the latent sources of both variance and co-variance obtained in the data and for interpreting the data scores (Joreskop and Sorbom 1993). The normality analysis was firstly performed in EFA. Skewness and Kurtosis values were determined as -1.511±.403 and 1.893 ±.788, respectively. The statistical value interval for 5% confidence interval of Skewness and Kurtosis values is expected to be ± 2.58. In addition, this range for 1% confidence interval is ± 1.96 (Liu et al., 2005). Kaiser Mayer Olkin's value (KMO) for the adequacy of the sample was found as .719. A high KMO value means that each variable in the scale can be perfectly predicted by other variables. Field (2000) also stated that 0.50 should be the lower limit for the KMO test and that the data set cannot be factored for KMO ≤ .50. Bartlett sphericity test was also statistically significant (χ^2 (210) = 477.701; p <.01). Ardingly the sample group is suitable for EFA analysis. Table 1 has shown EFA results.

Table 1. Exploratory Factor Analysis Results for Program Criteria

Criterion No.	Factor Load	Criterion No.	Factor Load	Criterion No.	Factor Load
C5	.850	C20	.715	C8	.605
C1	.817	C12	.702	C18	.602
C3	.807	C10	.700	C7	.576
C4	.770	C9	.699	C6	.572
C19	.753	C11	.683	C15	.505
C2	.753	C17	.634	C21	.457
C13	.737	C14	.627	C16	.454
Eigenvalues = 9.613, Announced Variance = 45,77 %					

From table 1 it can be claimed that the data is appropriate according to factor analysis results. The criteria have 45.77 % announced variance result under a single factor analysis. In addition, the factor analysis of each criteria ranger from .850 to .454 which means that the program evaluation survey has unidimensional. On the other hand, the reliability of the criterion form was provided with the Cronbach alpha coefficient, and the Cronbach alpha reliability coefficient for 21 criteria was calculated as .961. This reliability coefficient is predicted to be quite sufficient for the criterion form. Also, according to this result, it was seen that there was a high level of internal consistency among the criteria items. In addition, the Cronbach Alpha Coefficient is accepted as an indicator of the homogeneity of the feature studied. Accordingly, it can be said that the criterion form is homogeneous. There are different classifications in the literature for the interpretation of the Cronbach alpha reliability coefficient. According to the widely accepted approach, if the reliability coefficient alpha is greater than 9 (α ≥ .9), this is considered as a "perfect" (Cortina 1993; Streiner 2003; Tavakol & Dennick, 2011). As the Cronbach Alpha Coefficient approaches 1, the criterion form has a one-dimensional structure. Finally, the item statistics of the criteria items in the evaluation form were examined on the item-total correlation. Item total correlation is used to express the relationship between the score obtained from each criterion and the total score. It can be said from all results that the criterion form is dimensionless.

Table 2. Item-Total Statistics

Item	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
C1	51.50	84.35	.845	.904
C2	52.03	78.33	.706	.915
C3	50.94	89.35	.649	.907
C4	52.01	86.23	.764	.904
C5	52.12	91.76	.587	.925
C6	51.95	79.76	.716	.925
C7	51.44	86.23	.695	.951
C8	50.86	94.49	.740	.958
C9	51.50	86.78	.780	.904
C10	52.02	86.01	.726	.950
C11	50.94	88.78	.601	.949
C12	50.85	86.34	.757	.897
C13	52.10	94.45	.680	.900
C14	52.64	98.65	.535	.924
C15	51.31	87.38	.671	.896
C16	52.07	86.99	.684	.900
C17	50.95	82.24	.793	.904
C18	51.29	90.10	.517	.955
C19	51,55	87.19	.632	.945
C20	52,04	85.65	.855	.950
C21	52,13	83.34	.796	.930

Table 2 shows how the Cronbach alpha value changes with the criterion item after removing undesirable items. Cronbach alpha if item deleted column indicated that the lowest score is .896 (higher than .80) therefore reliability coefficient criteria has met. Accordingly, criterion survey prepared with these 21 items.

b) Data-Model Fit

This study was used unexpected value or, standardized residual value (StRes) so that comparing whether data-model fit is suitable or not. Linacre (2014) claimed that for appropriate data and model, less than 1 % of StRes value should be located in the range of ±3. Similarly, less than 5 % of StRes value should be located in the range of ±2. In this study, outside the range of StRes value of ±2 was 2.3 %; and, outside the range of StRes value of ±3 was .4 % which means that data-fit model assumption was met by considering StRes value.

c) Local Independence

Local independence is related with unidimensionality, but it demonstrates the relationship between responses of criteria survey and response of each item. Local independence supposes that if unidimensionality is provided, local independence is provided too. Therefore, there is no need for extra test for local independence.

Significance of the Study

In the literature, there is limited study on evaluations of science teacher curriculum, especially from science teacher educators' perspective. All undergraduate science teacher curriculum in the universities use CoHE's (2018) teacher curriculum. Science teacher educators taught compulsory and elective lessons prepared by CoHE (2018). For this reason, science teacher educators' evaluations are significant. Therefore, this study aims to investigate science teacher educators' evaluations about science teacher curriculum which renewed in 2018. These evaluations include content, related activities, timing and, assessment and evaluation about the courses. By considering this aim, these research questions revealed;

In the literature, there is limited study on evaluations of science teacher curriculum, especially from science teacher educators' perspective. All undergraduate science teacher curriculum in the universities use CoHE's (2018) teacher curriculum. Science teacher educators taught compulsory and elective lessons prepared by CoHE (2018). For this reason, science teacher educators' evaluations are significant. Therefore, this study aims to investigate science

teacher educators' evaluations of the science teacher curriculum which were renewed in 2018. These evaluations include content, related activities, timing and, assessment and evaluation of the courses.

By considering this aim, these research questions revealed;

1. What is the distribution of the jury-criteria item calibration map of the science teacher curriculum?
2. How are jury members generous/ungenerous behavior while evaluating the science teacher curriculum?
3. How are statistics of analysis of each criterion used in evaluating the science teacher curriculum?

Methodology

Research Design

This is a cross-sectional and particular scanning model approach to evaluate science teacher curriculum (Creswell, 2002). In the model, data are obtained by only one specific test, form, or survey without interfering with the existing situation (Fraenkel & Wallen, 2006). The purpose of this design is to explain a current situation by analyzing and describing it (Gay et al., 2009). For this study, science teacher curriculum of the 2018 is selected to evaluation of a single measurement.

Participants

From purposeful sampling methods, criterion sampling strategy used while selecting participants for this study (Sandelowski, 2000). At first, universities who applied CoHE 2018 science teacher curriculum (N=33) is selected. It is found that in these universities there are 138 academicians who worked at science teacher education department. Then researchers who did not have doctorate (research and teaching assistants) are eliminated since they have fewer experiences in science teacher curriculum. In addition, researchers realized that even though some academicians worked at science teacher education department, they did not have doctorate in science teacher education. Therefore, there are excluded in the study and remained participant number is fifty-seven. These 57 academicians who have doctorate in science teaching and work at science teacher education department is the population of this study. Researchers send an e-mail to all determined participants however, only 34 of them replied and participated. Sampling error was found 9,04 % according to the Salant and Delman's (1994) sampling error formula. In addition, reliability of the sampling is found 90% which is higher four small samplings. Eventually, participants of this study are 34 (23 of them were female and 11 of them were male) science teacher educators. They have on average 12.2 years experience in science teaching (SD=5.2 within an experience range of 1-24 years). For this study each science teacher educators were coded as numbers which create "Jury" and coded as J1, J2, J3, …, J34.

Data Collection

For this study, data collection tool is evaluation form for science teacher curriculum. This form is prepared by researchers of this study by considering previous studies' teacher curriculum criteria (Juttner et al., 2013; Kahle et al., 2000; You, 2016). and four dimensions of the curriculum which are 1. aims, goals, and objectives, 2. subject matter, 3. learning experiences, and 4. evaluating approaches (Ornstein & Hunkins 2009). While developing evaluation form, higher content validity is needed. The most common way to ensure content validity is to set up a subject expert panel that determines the importance of items on a scale. Quantitative and qualitative indicators obtained from the examination of the items planned to be included in the scale by experts for content validity can be useful in identifying the wrong steps and corrected content during the scale development phase. It is essential to use a quantitative criterion when estimating content validity. These criteria used by experts in content validity are Content Validity Index (CVI) and Content Validity Ratio (CVR). On the one hand, content validity ratio is an internationally accepted criterion for deciding whether each item will be included in the scale or not. On the other hand, the content validity index is the average CVR for all items in the final scale. In other words, CVR is used to determine whether each item is necessary and CVI is used to determine the relationship of each item in the scale with the scale used. The CVI is calculated by using the degree of agreement of the experts on the relevance and clarity of the items.

Accordingly, the construct validity of the Science Curriculum Evaluation Form (SCEF) was carried out in 6 steps defined by Polit and Beck (2006). These are 1-Preparing a content verification form, 2-Selecting a review panel of experts, 3- Performing content verification, 4-Examination of the area and elements, 5- Providing points for each item, and 6- Calculation of CVR, I-CVI and S-CVI. CVR value was calculated according to the Lawshe (1975) and Ayre and Scally (2014) formula and CVI value was calculated the recommendations reported by Lynn (1986) and Polit and Beck (2006). Based on the relevant literature, a 24-item SCEF was prepared to meet the expectations of the commission members. SCEF was submitted to the approval of an expert commission of 14 people, consisting of a linguistics expert, an assessment and evaluation expert, and twelve science education department faculty members (4 Professors, 5 Associate Professors, 3 Assistant Professors) by convenient sampling method.

Development of Science Curriculum Evaluation Form

The development process of the science curriculum evaluation form **(**SCEF) form was started with the calculation of CVR values, which were suggested by Lawshe for the first time and were an indicator of its structural validity. However, the arguments suggested by Ayre and Scally (2014) were used in the interpretation of the CVR values. Here it was calculated according to the equation CVR=A/(N/2)-1. Where N: the total number of experts, A: the number of experts who rated "relevant" (those who gave 3 or 4 points). According to Ayre and Scally (2014), CVR can be used as a statistical tool used to accept or reject certain substances. The number of experts who gave 3 or 4 points to the criteria form was considered in the calculations. In addition, "What is your suggestion?" from the experts who marked the "need to be fixed" option; Experts who marked the "must be removed" option were asked to give a second opinion as "Why?" In the interpretation of CVR values, Ayre and Scally's (2014) proposed content validity criterion ($CVR_{critical}$=critical CVR) for each item with a positive value at $\alpha=0.05$ significance level was examined.

According to Ayre and Scally, CVR=$CVR_{critical}$ value is a value needed to eliminate the chance of being called "appropriate" for each item in the scale and to decide whether an item is suitable. According to the evaluations of 14 experts, the $CVR_{critical}$ value recommended by Ayre and Scally (2014) is .51. Accordingly, it was determined that the CVR values determined for each item of the FMDF form for the number of fourteen experts at the $\alpha=0.05$ significance level was greater than the recommended $CVR_{critical}$ value for all.

In addition, as Ayre and Scally (2014) stated, at least 11 people in the said commission are capable of adjudicating on the articles. From all these results, statistical significance was found for each item in the criterion form. On the other hand, since the CVR statement, which was previously suggested by Lawshe and made some corrections by Ayre and Scally (2014), is based on an empirical approach. Accordingly, whether each item in the criteria form would be used as a criterion was determined by the content validity ratio, I-CVR. In addition, the S-CVI value was calculated to determine whether there was agreement among the experts. There are two separate CVI forms that represent the CVI. These are the I-CVI values that define the item coverage index and the S-CVI values that indicate the overall content validity of the scale. In addition, S-CVI can be calculated by two methods. In the first of these, the average of the I-CVI scores of all items in the scale is found as S-CVI/Ave. In the other, the ratio of experts who marked the relevance of the items in the scale as 3 or 4 gives S-CVI /UA. S-CVI /UA is called universal-based agreement method-scale level content validity index. These concepts have been previously discussed in Lynn (1986), Davis (1992), and Polit and Beck (2006). It has also been suggested by According to the recommendations; the minimum value of I-CVI should be 0.78 or greater in studies consisting of 5 or more experts (Orts-Cortés, 2013).

After these calculations, scores from the FMDF form were converted to kappa values to account for the chance factor among participants, and a modified Kappa index was used to estimate I-CVI [Wynd CA et al, 2003]. Modified Kappa (k*) is an index of agreement among experts that indicates that the item is more than likely to be a feature other than being relevant, clear, or interesting (the degree of agreement beyond chance) [Wynd CA at all, 2003]. However, the modified kappa sequence suggested by Fleiss (1971) was used to evaluate the Kappa value. Accordingly, the rating scale for Fleiss kappa was “excellent (≥0.74)”, “good (0.60 to 0.73”), “moderate (0.40 to 0.59”) and “poor (≤0.39)”. as recommends. Since the kappa values of all the items in the PDF form were above ≥0.74, the degree of agreement between the participants was evaluated as “excellent”. Accordingly, no potentially problematic items were found in the form. The equations used to calculate the kappan are as follows. pc=[N!/A!(NA)!] 〖0.5〗^N and k=(I-CVI-pc)/(1-pc) where k: Modified kappa coefficient, pc: probability of random correlation coefficient (chance-congruence ratio), N: number of experts, A: number of experts who rated “relevant” (those who gave a score of 3 or 4). Microsoft Excel 2007 software program was used in all calculations. From all these results, the content validity of SCEF was found to be statistically significant. Thus, SCEF consisting of 34 items in 5-likert type, was prepared between the options 'not suitable' corresponding to 1 point and 'completely suitable' corresponding to 5 points in the criterion form.

Table 3 indicates calculations for CVI and CVI values for SCEF.

Data Analysis

Data analysis is conducted by using MFRM frame with FACETS program which is developed by Linacre (2014). Previously, many of the educators, uses parametric statistic tests so that analyzing their data. However, multiple-choice test data are not always meet the criteria of parametric assumptions because of the facts there is no agreement on what causes slight deviation from an assumption (Siegel, 1956). To solve this problem Rasch (1960) suggested theory-based, informative, valid and reliable solutions for science educators Sondergeld and Johnson (2014). Boone et al. (2011) claimed that Rasch Analysis has a strong quantitative approach, however it should be used while a research problem needs qualitative analysis and approach. Since the problem statement of this study is appropriate for nature of Rasch analysis, it is used. In this study two-facet Rasch Model design was used to analyze jury members evaluations about criteria. Accordingly, for both facet score of criteria and jury members are calculated, independently.

Table 3. Evaluation for Science Curriculum Evaluation Form (SCEF)

Dimension	Item	Expert														Score				N_A	CVI	UA	CVR	pc $x10^{-3}$	k*	Rating[a]
		E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**E11**	**E12**	**E13**	**E14**	**4**	**3**	**2**	**1**							
Aims, Goals and Objectives	Crt1	4	4	**3**	4	4	4	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt2	4	**3**	4	4	4	4	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt3	4	4	4	4	4	4	4	4	**3**	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt4	4	4	4	4	4	**3**	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt5	4	4	4	4	4	4	4	4	4	**3**	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
Subject Matter	Crt6	4	4	4	**2**	4	4	4	4	4	4	4	**3**	4	4	12	1	1		13	.92	0	.85	.85	.92	Excellent
	Crt7	4	4	4	4	**3**	4	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt8	4	4	4	4	4	4	**3**	4	4	4	**2**	4	4	4	12	1	1		13	.92	0	.85	.85	.92	Excellent
	Crt9	**3**	4	4	4	4	4	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt10	4	4	4	4	4	4	4	4	4	**3**	4	4	**2**	4	12	1	1		13	.92	0	.85	.85	.92	Excellent
	Crt11	4	4	4	4	4	4	4	**3**	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
Learning Experiences	Crt12	4	4	4	4	4	4	4	4	4	4	4	4	4	**3**	13	1			14	1	1	1	.061	1	Excellent
	Crt13	4	4	4	4	4	4	4	4	4	4	4	3	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt14	4	4	4	4	4	**3**	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt15	4	4	4	4	4	4	4	4	**2**	4	4	**3**	4	4	12	1	1		13	.92	0	.85	.85	.92	Excellent
	Crt16	**3**	4	4	4	4	4	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt17	4	4	4	4	4	4	4	4	4	4	4	4	4	**3**	13	1			14	1	1	1	.061	1	Excellent
	Crt18	4	4	**3**	4	4	4	4	4	4	4	4	4	4	4	13	1			13	.92	1	.85	.85	.92	Excellent
Evaluating Approaches	Crt19	4	4	4	4	4	4	**3**	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt20	4	**3**	4	4	4	4	4	4	4	4	4	4	4	4	13	1			14	1	1	1	.061	1	Excellent
	Crt21	4	4	4	4	**2**	4	4	4	4	4	4	**3**	4	4	12	1	1		13	.92	0	.85	.85	.92	Excellent
Proportion relevant		**1.0**	**1.0**	**1.0**	**.97**	**.97**	**1.0**	**1.0**	**1.0**	**.97**	**1.0**	**.97**	**1.0**	**.97**	**1.0**	**S-CVI/Ave**					**.98**					
Average proportion of items evaluated by 14 experts, S-CVI/Ave*															**.98**	**S-CVI/UA**						**.82**				

* NA: Number of Agreement, there is no CVR $_{critical}$ (.571)value according to Ayre and Scally (2014), I-CVI: Content Validity; Pc: probability of random compromise; k*: kappa coefficint, k* values: poor ≤0.39, weak = 0.40–0.59; good = 0.60–0.73; excellent ≥0,74 (Fleiss, 1971), S-CVI/Ave* (based on proportion relevance): Average proportion of “relevant” scores through experts, S-CVI/Ave (based on I-CVI): mean I-CVI scores of all items

Results

In this section, results of each research question are given respectively. All of the analyses are conducted with MFRA. Results from analyses MFRA with their interpretation can be found in sub-sections.

Results of Jury-Item Calibration Map

The first research question was "What is the distribution of jury-criteria item through calibration map of science teacher curriculum?". On the data calibration map, each jury member and criteria's rating scores are demonstrated. In the map, science teacher curriculum is coded as "prgm" and evaluated by 34 Jury by considering 21 criteria. The data calibration map of the MFRA statistics depends on this study's data was showed in Figure 1.

```
+----------------------------------------------------------+
|Measr|PRGM  |JURY              |CRITERIA          |RATIN|
|-----+------+------------------+------------------+-----|
|   3 +      + J11              +                  + (5) |
|     |      |                  |                  |     |
|     |      |                  |                  |     |
|     |      | J2               |                  |     |
|     |      |                  |                  |     |
|     |      | J1    J29   J3   |                  |  4  |
|     |      | J31              |                  |     |
|   2 +      + J19   J5    J6   +                  +     |
|     |      |                  |                  |     |
|     |      | J26              |                  |     |
|     |      | J10   J23   J9   |                  |     |
|     |      |                  |                  |     |
|     |      | J18   J21        |                  |     |
|     |      |                  |                  |     |
|     |      | J25              |                  |     |
|     |      | J27              |                  |     |
|   1 +      +                  +                  +     |
|     |      |                  |                  | --- |
|     |      | J16              | C20              |     |
|     |      | J22   J24   J28  |                  |     |
|     |      |                  |                  |     |
|     |      | J33              | C13  C17  C7     |     |
|     |      | J30   J7         | C12  C3          |     |
|     |      | J8               | C15  C18         |     |
|     |      |                  | C10  C19  C21    |     |
*   0 * P1   *                  * C1   C14  C4     *     *
|     |      |                  |                  |     |
|     |      |                  | C16  C5          |     |
|     |      |                  |                  |  3  |
|     |      | J14              | C2               |     |
|     |      | J32              | C11  C6          |     |
|     |      | J13   J15        |                  |     |
|     |      | J4               | C8               |     |
|     |      | J12              | C9               |     |
|     |      |                  |                  |     |
|  -1 +      +                  +                  +     |
|     |      |                  |                  |     |
|     |      |                  |                  | --- |
|     |      |                  |                  |     |
|     |      | J17              |                  |     |
|     |      |                  |                  |     |
|     |      |                  |                  |     |
|     |      | J20              |                  |     |
|  -2 +      +                  +                  + (1) |
|-----+------+------------------+------------------+-----|
|Measr|+PGRM |+JURY             |-CRITERIA         |RATIN|
+----------------------------------------------------------+
```

Figure 1. Data Calibration Map

As stated Figure1, juries' evaluation scores of the program vary from 1 to 5, and average criteria score is around 3. While Jury11 (8.22 logit) and Jury2 (2.61 logit) are the most generous members during evaluating of the program, Jury20 (-1.92 logit) is the least generous one. Since most of the jury members are located at the rating scale of around three, it indicated that they feel neutral about improvement of new science teacher curriculum.

Furthermore, C20 has the highest rating score (0.72 logit) and C9 has the lowest rating score (-0.80 logit). This result has demonstrated that scores about rating scale of 3, which means that criteria developed for science teacher curriculum is uniformly distributed among each criterion.

Results of Each Jury's Generosity and Ungenerosity Behavior

Second research question was *"How is jury's generosity and ungenerosity behavior while evaluating science teacher curriculum?"*. To answer this question each criterion for used in evaluating the program regarding the logit values for the judge facets is examined. Finding facet statistics have given in Table 4 which shows jury members' evaluation of the criteria.

Table 4. Measurement report on the generosity and ungenerosity behaviors of jury

Jury Member	Observed Average	Fair Average	Model		Infit		Outfit	
			Measure	Error	Square Average	Z	Square Average	Z
J11	5.00	4.99	8.22	1.84	max			
J2	4.10	4.09	2.61	.41	2.12	2.5	2.13	2.5
J1	4.00	4.00	2.28	.40	.65	-.9	.65	-.9
J3	4.00	4.00	2.28	.40	2.16	2.5	2.17	2.5
J29	4.00	4.00	2.28	.40	.95	.0	.95	.0
J31	3.95	3.95	2.12	.40	1.48	1.2	1.50	1.2
J5	3.90	3.91	1.97	.39	.49	1.06	.43	-1.8
J6	3.90	3.91	1.97	.39	.36	-2.2	.37	-2.1
J19	3.90	3.91	1.97	.39	1.38	1.0	1.41	1.1
J26	3.86	3.86	1.81	.39	.92	-.1	.90	-.1
J9	3.81	3.81	1.66	.38	.41	-2.0	.39	-2.1
J10	3.81	3.81	1.66	1.62	1.5	-164	1.64	1.6
J23	3.81	3.81	1.66	.38	.37	-2.2	.35	-2.3
J18	3.71	3.72	1.38	.37	1.03	.2	1.00	.1
J21	3.71	3.72	1.38	.37	.92	-.1	.92	-.1
J25	3.67	3.67	1.25	.26	.65	-1.1	.67	-1.0
J27	3.62	3.63	1.12	.36	1.35	1.0	1.32	.9
J16	3.48	3.48	.75	.34	1.65	1.9	1.70	2.0
J22	3.43	3.44	.63	.34	.69	-1.0	.70	-1.0
J24	3.43	3.44	.63	.34	1.97	2.6	2.03	2.7
J28	3.43	3.44	.63	.34	1.01	.1	.99	.0
J33	3.33	3.34	.41	.33	1.52	1.6	1.53	1.6
J7	3.29	3.29	.30	.33	1.34	1.1	1.34	1.1
J30	3.29	3.29	.30	.33	1.34	1.1	1.34	1.1
J8	3.24	3.24	.20	.32	.97	.0	.99	.0
J14	2.95	2.96	-.41	.31	.85	-.4	.84	-.4
J32	2.90	2.91	-.50	.31	.23	-3.9	.22	-3.9
J13	2.86	2.86	-2.11	.82	.59	-.6	.58	-.6
J15	2.86	2.86	-.60	.31	.63	-1.3	.64	-1.3
J4	2.81	2.81	-.69	.31	1.17	.6	1.15	.6
J12	2.76	2.76	-.79	.31	.65	-1.2	.65	-1.2
J17	2.38	2.38	-1.54	.31	.58	-1.6	.58	-1.6
J20	2.19	2.19	-1.92	.31	1.07	.3	1.06	.2
Standard Deviation	3.50	3.50	1.04	.40	1.02	-1.2	1.02	-.1

Model, Sample: RMSE .47 Adj (True) S.D. 1.66 Separation 3.51 Strata 5.01 Reliability (not inter-rater) .92

Model, Fixed (all same) chi-square: 400.8 d.f.: 32 significance (probability): .00

Model, Random (normal) chi-square: 22.9 d.f.: 31 significance (probability): .85

Table 4 demonstrated logit value of each jury member, input value, and outfit value which creating reliability of each facet. RMSE value found .47 which is smaller than critical 1.00 value. In addition, high reliability index demonstrated the difference is reliable (Haiyang 2010). In addition, chi square results and separation index compares whether there is statistically significant difference or not. Table 4 showed that while separation index is calculated as 4.10 reliability index value was calculated .85 ($\chi 2=72.2$, $p<.05$) which means that there is statistically significant difference among jury members' evaluation of the program. Therefore, null hypothesis which is rejected in terms of the generosity and ungenerosity behavior of jury members' evaluation. While J11 is the most generous one (5.00 average point out of 5.00), J20 is the most ungenerous (2.19 average point out of 5.00).

Results of Analysis of Criteria of Science Teacher Curriculum

Third research question was *"How is statistics of analysis of each criterion used in evaluating the science teacher curriculum?"*. To answer this question each criterion for used in evaluating the program regarding the logit values for the judge facets is examined. Finding facet statistics have given in Table 5 which shows average results of each criterion.

Table 5. The measurement report results for evaluation criteria of undergraduate program

Criteria	Observed Average	Fair Average	Model		Infit		Outfit	
			Measure	Error	Square Average	Z	Square Average	Z
C20	3.21	3.20	.72	.27	.77	-.9	.78	-.8
C7	3.33	3.33	.43	.27	1.20	.8	1.25	.9
C13	3.33	3.33	.43	.27	1.10	.4	1.00	.0
C17	3.33	3.33	.43	.27	.97	.0	1.03	.2
C3	3.39	3.40	.28	.28	1.05	.2	1.08	.3
C12	3.39	3.40	.28	.28	.93	-.1	1.02	.1
C15	3.42	3.43	.20	.28	1.21	.8	1.23	.8
C18	3.42	3.43	.20	.28	1.32	1.2	1.44	1.5
C10	3.45	3.46	.12	.28	.97	.0	.90	-.3
C19	3.45	3.46	.12	.28	.65	-1.4	.66	-1.3
C21	3.45	3.46	.12	.28	1.06	.3	1.19	.7
C1	3.52	3.53	-.04	.28	1.12	.5	1.08	.3
C4	3.52	3.53	-.04	.28	.90	-.3	.87	-.4
C14	3.52	3.53	-.04	.28	.91	-2	.94	-.1
C5	3.58	3.59	-.20	.29	.82	-.6	.91	-.2
C16	3.58	3.59	-.20	.29	.98	.0	.96	-.0
C2	3.64	3.65	-37	.29	1.05	.2	1.04	.2
C6	3.67	3.68	-.45	.29	1.03	.1	1.13	.5
C11	3.67	2.68	-.45	.29	.67	-1.3	.64	-1.4
C8	3.76	3.77	-.71	.30	1.54	1.8	1.59	1.9
C9	3.79	3.80	-.80	.30	.75	-.9	.75	-.9
Standard Deviation	3.50	3.50	.00	.28	1.00	.0	1.02	.1

Model, Sample: RMSE .28 Adj (True) S.D. .26 Separation .93 Strata 1.57 Reliability .46

Model, Fixed (all same) chi-square: 38.3 d.f.: 20 significance (probability): .01

Model, Random (normal) chi-square: 13.5 d.f.: 19 significance (probability): .81

According to Table 5, C9 (Content and objectives in the program are interrelated.) took the highest average point (3.79) from jury members whereas C20 (Content is responsive to the individual needs of students.) took the lowest average point (3.21). Reliability index of the results has found .46 which is significantly smaller than .80 and means that some jury might have bias on evaluating some criterion. Accordingly, deviation of results is found .26 (<1.00), there are it is necessity to look chi square and significance values. According to the statistics results chi square value indicated difference among results are meaningful ($\chi 2=38.3$, sd=20, p<.01). Therefore, null hypothesis is rejected, and it can be claimed that there are statistically significant difference values of criteria that evaluating the program. However, for this study bias means that there are unexpected choices while evaluating the science teacher curriculum. Table 6 indicated that which jury had bias on evaluating which criteria.

Table 6. Jury's Bias on criterion while evaluating the science teacher curriculum

Score	Exp.	Resd	StRes	Jury	Criteria
1	3.6	-2.6	-4.1	J27	C15
2	4.0	-2.0	-3.8	J2	C12
2	4.0	-2.0	-3.8	J2	C8
2	3.9	-1.9	-3.4	J19	C21

Table 6. indicated that J27 had negative bias on the C15 (Learning activities in the program are teacher-centered). While J27's average score is 3.6, 1 point had gain to C15. This indicated that J27 did not think that science teacher curriculum is teacher-centered. Similarly, J2 had bias on C12 (Content provides an enjoyable environment to students) and C8 (Time is not enough to teach knowledge and skills in content.) This result indicated that J2 thought that time is enough for teaching knowledge and content however these are not enjoyable. Lastly, J19 had negative bias on C21 (The activities in the program content are boring). This result indicated that according to the J19's point of view activities was not boring.

Discussion

In this present study, the science teacher curriculum updated in 2018 in Türkiye was evaluated by considering various criteria according to MFRM (Many-Facet Rasch Model). According to this analysis, each criteria's consistency and evaluations of each criterion are examined. In addition to that, science teacher curriculum academicians' (juries') generosity and ungenerosity behavior during the evaluation of the program is analyzed. Lastly, whether there is rater bias among jury members is analyzed.

At a first glance, result demonstrated that academicians have neutral evaluations about updating the 2018 science teacher curriculum. *Evaluating approaches* dimension has the highest scores from academicians which means that they agree that updating science teacher curriculum allow formative, summative and authentic assessment. Similarly, Veal (2004) argued the importance of both formative and authentic assessment. According to the results of this study, science teacher educators thought that the program is suitable for different assessment strategies.

On the other hand, the *subject matter* dimension has the lowest average score. This result indicates that the science teacher curriculum subject matter is not much understandable, and interesting. In addition to that, concrete examples are not much sufficient. Science teacher curriculum' subjects should be related to the middle school science curriculum and both of them should be up to date. For this reason, they are improving by considering students', teachers' and society's needs and expectations. However, in this study jury members gave low average points to criteria which focus on the compatibleness of science teacher curriculum to middle school curriculum. There are compulsory courses in the middle school science curriculum (CoHE, 2018). In addition, Cronin-Jones (1991) stated the importance of how to teach concepts to little students. During science teacher curriculum preservice science teachers take both discipline specific courses and pedagogy courses. While they learn discipline-specific knowledge in discipline courses, they learn how to teach them in pedagogy courses. Accordingly, in science teaching courses they experience micro-teach of these contents and concepts for middle schoolers. This nature of the science teacher curriculum support raising effective teachers. Juries who evaluated the program also gave higher points to these criteria.

Results about the juries' generosity and ungenerosity behaviour indicated that there are science teacher educators whose have generous and ungenerous characteristics during evaluating the program. While J11 was the most generous one with 5.00 average point, J20 was the most ungenerous one with the 2.19 average point. Then whether they are bias of jury members for each criterion is examined. Previous studies conducted on MFRM (Boone et al., 2011; Juttner et al., 2013) have stated that rater bias should need to considerate since it is affected reliability and validity scores. Similar to previous studies in this study bias has found on some jury's scoring behavior. However, reliability of the scoring has found 0.92 which indicated highest reliability of the juries. They are reliably ranked in terms of generosity and ungenerosity behaviour and differ from each other. Farrokhi and colleagues' (2012) study also found that jury members might have generosity or ungenerosity behavior while scoring. If the total reliability is higher than .80, it is normal to have generosity or ungenerosity behavior while evaluating programs, projects, or curriculums. Finally, all this result supported that the MFRM can be used as an alternative measurement model in evaluating the curriculums or programs. In addition, developed surveys or forms can be use in parts since they have multiple dimensions. For this study, fit results and person and item reliability scores support to use this item in the future while evaluating curriculums or programs in teacher education.

Recommendations for Further Research

There are recommendations based on this study's results and future studies. First, MFRA is a strong quantitative statistic, qualitative interviews can support the results of the statistics. In other worlds, adding a in qualitative part to the study might have strength in interpreting results and the study itself. Second, in this study, only science teacher educators created a jury in order to evaluate the science teacher curriculum. Science teacher curriculum developers or preservice science teachers can be added to the study in order to add additional perspectives to the evaluation. These additional perspectives might appear on how developers and preservice science teachers think about the science teacher curriculum. Lastly, before evaluating the program, short training sections about the evaluation of the program can be given to all juries to eliminate possible bias.

Contribution Rate of the Researchers

All authors contributed to the manuscript equally.

Financial Support and Acknowledgment

The authors declared that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors have disclosed no conflict of interest.

References

Aikenhead, G. S. (1997). Toward a First Nations cross-cultural science and technology curriculum. *Science Education*, *81*(2), 217.

Atkin, J. M. (1998). The OECD study of innovations in science, mathematics and technology education. *Journal of Curriculum Studies*, *30*(6), 647.

Ayre, C., & Scally A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47*(1), 79.

Bailes, L. P., & Nandakumar, R. (2020). Get the most from your survey: An application of Rasch analysis for education leaders. *International Journal of Education Policy and Leadership*, *16*(2), n2.

Bawane, J., & Spector, J. M. (2009). Prioritization of online instructor roles: implications for competency-based teacher education programs. *Distance Education*, *30*(3), 383.

Bencze, L., & Carter, L. (2011). Globalizing students acting for the common good. *Journal of Research in Science Teaching*, *48*(6), 648.

Bencze, L., & Hodson, D. (1999). Changing practice by changing practice: Toward more authentic science and science curriculum development. *Journal of Research in Science Teaching*, *36*(5), 521.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253.

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, *95*(2), 258.

Council of Higher Education (CoHE). (2007). *Teacher Education and Faculty of Education (1982-2007).* CoHE.

Council of Higher Education (CoHE). (2018). *Teaching Science Programs.* CoHE.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall.

Cronin-Jones, L. L. (1991). Science teacher beliefs and their influence on curriculum implementation: Two case studies. *Journal of Research in Science Teaching*, *28*(3), 235.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76(*5), 378.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2009). *Educational research competencies for analysis and applications*. Merrill/Pearson.

Juttner, M., Boone, W., Park, S., & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability*, *25*(1), 45.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology, 28*(4), 563.

Linacre, J. M. (2012). Many-facet Rasch measurement: Facets tutorial. Retrieved April 24, 2017 from http://www.winsteps.com/a/ftutorial2.pdf

Lynn M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382.

Okcabol, R. (2005*). Teacher training system: historical development, current situation, and a systems approach to the problem of teacher education*. Utopya Publishing.

Olivia, P. F. (1997). *Developing the curriculum* (4th ed.). Longman.

Ornstein, A. C., &Hunkins, F. P. (2009).*Curriculum foundations, principles and issues* (6th ed.). Pearson Education.

Orts-Cortés, M. I., Moreno-Casbas, T., Squires, A., Fuentelsaz-Gallego, C., Maciá-Soler, L., & González-María, E. (2013). Content validity of the Spanish version of the Practice Environment Scale of the Nursing Work Index. *Applied Nursing Research*, *26*(4), e5.

Oon, P. T., & Fan, X. (2017). Rasch analysis for psychometric improvement of science attitude rating scales. *International Journal of Science Education*, *39*(6), 683.

Orhan, E. E., (2017). What teachers think about teacher education they received in Turkey? A qualitative research. *Education and Science, 42*(189), 197.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health*, *29*(5), 489.

Popkewitz, T. S. (1994). Professionalization in teaching and teacher education: Some notes on its history, ideology, and potential. *Teaching and Teacher Education*, *10*(1), 1.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Robinson, B., & Latchem, C. (2003). Teacher education: Challenge and change. In B. Robinson & C. Latchem (Eds.), *Teacher education through open and distance learning* (pp. 1-27). Routledge.

Sandelowski, M. (2000). Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in Nursing & Health*, *23*(3), 246.

Siegel, S. (1957). Nonparametric statistics. *The American Statistician, 11*(3), 13.

Sondergeld, T. A., & Johnson, C. C. (2014). Using Rasch measurement for the development and use of affective assessments in science education research. *Science Education*, *98*(4), 581.

Unal, S., Çoştu, B., & Karataş, F. Ö. (2004). Program development activities for science education: An overview. *Journal of Gazi University Faculty of Education, 24*(2), 183.

Veal, W. R. (2004). Beliefs and knowledge in chemistry teacher development. *International Journal of Science Education*, *26*(3), 329.

Veneziano L. ve Hooper J. (1997). A method for quantifying content validity of health-related questionnaires. *American Journal of Health Behavior, 21*(1), 67.

Wei, B. (2020). The change in the intended Senior High School Chemistry Curriculum in China: focus on intellectual demands. *Chemistry Education Research and Practice*, *21*(1), 14.

You, H. S. (2016). Rasch validation of a measure of reform-oriented science teaching practices. *Journal of Science Teacher Education*, *27*(4), 373.