Multifaceted Rasch Analysis for Clinical Supervision instrument assessment of Islamic Religious Education Teachers

Yuni Asdhiani[1*], Ari Saptono [2] , Komarudin [3]

[1]Supervisor of Islamic Religious Education at Junior High Schools Bekasi , Indonesia

[2, 3]State University of Jakarta, Indonesia

* pengawasyasdhiani@gmail.com

ABSTRACT

This study aims to analyze (1) validation of clinical supervision assessment instruments; (2) the reliability of the clinical supervision assessment instrument. This study was adapted using research & development using the borg & gall development model. The results of the expert assessment were analyzed with the help of the Multi-Rater Facet Rasch (MRFR) using the facet software. The results of this study indicate: (1) validation results The content of the clinical supervision instrument for Islamic Religious Education teachers is based on the CVR value of 0.99, so all items are declared valid and suitable for further research (3) The reliability results in large-scale research with a reliability value of 0.98 means that the instrument has high quality stability. high enough it shows good inter-rater reliability.

Keywords

Clinical supervision, Assessment, Multi Facet Rasch Model

Article Received: 18 October 2020, Revised: 3 November 2020, Accepted: 24 December 2020

Introduction

Supervision activities are carried out in an effort to improve and control all activities so that they are carried out optimally. Supervision / supervision is ensuring activities are effectively carried out by those who are responsible for carrying them out. Supervisors are usually those who focus on the day-to-day activities of the department and evaluate those who perform them. The most effective managers and supervisors are also leaders (Kären Matison Hess 2012). Clinical supervision is a special practice at the highest level in social work practice because clinical supervisors prepare social workers to train independently without the need for further supervision (Linda Openshaw 2012). The definition of clinical supervision is adopted from the medical profession, namely the process of developing the skills and knowledge of training participants in practical activities.

Supervision education activities are carried out in order to improve learning activities. In general, the current supervision appears on the initiative of the principal or supervisor, not from the teacher. Ideally, efforts to improve learning activities come from the teacher concerned, not from other parties. Initiatives to improve teaching abilities come from the teacher itself, it is very important to be developed which is the seed that will soon be developed into awareness-based teacher capacity building activities, which is the core of the concept of clinical supervision. The purpose of clinical supervision is to help teachers develop and improve their professionalism through joint planning (teacher and supervisor), observation and feedback (Gursoy et al. 2013)

Differences were also found between the views of supervisors and supervisors in the clinical setting. Teachers want their teaching stage to be taken into account by supervisors when choosing a method of supervision, and they are also expected to be treated the same throughout their clinical practice, whereas supervisors have not yet assessed this supervisory behavior.

Supervision activities are the main task and function of school principals and supervisors, supervision activities aim to improve learning activities, in this condition the teacher appears passive. Efforts to improve learning activities should ideally grow from the teachers themselves, not from the school principal or regional supervisors. Clinical supervision is an effort made to improve student learning abilities and teacher teaching abilities, with several stages.

Effective and professional supervision is an activity towards quality education, the quality of student learning is directly influenced by the quality of teaching in the classroom, and quality education depends on how well teachers are trained and supervised, by helping teachers improve the skills and competencies required in planning, delivery of lessons and learning strategies, and increasing teacher effectiveness in learning so that it can be seen in student learning outcomes. A quality learning process and good student learning outcomes are an indicator of the success of the school supervisor's performance (Fteiha and Abdawi 2017; Kemendikbud 2016)

The instructional competence of teachers is still considered low to date so that the impact of producing student academic achievement is not as expected here the need for supervisory action, which involves ensuring the teacher fulfills instructional responsibilities effectively and efficiently, the teacher must demonstrate high academic standards through periodic checks to improve quality their work. School supervisors have a strategic role in improving the quality of education, referring to the responsibility to provide services and assistance to teachers and school principals that will influence the improvement of the educational process and outcomes (Okorji and Ogbo 2013).

Supervision or supervision is needed to improve the quality of education, without an effective supervisory program and professional educational programs cannot be controlled properly, supervision is likened to the glue for the success of educational programs, supervision has become an integral component in the process of school operations, and can function as evaluations that measure and access teaching effectiveness.

One of the supervisory models that can help improve teacher professionalism is through clinical supervision. Clinical supervision is the process of facilitating the professional growth of a teacher (Meriza 2018; Okorji and Ogbo 2013; Olibie and Ezeoba 2016; Zahraini, Situmorang, and Siburian 2018), especially by observing teacher instructional practices, providing teachers with feedback about classroom interactions and helping teachers utilize that feedback to make teaching more effective (Olibie and Ezeoba 2016), and improving teacher performance (Bessong and Ojong 2010; Chidi and Victor 2017; Sarfo and Cudjoe 2016; Veloo, Komuji, and Khalid 2013)

Clinical supervision is a type of surveillance that satisfies the requirements of good supervision in the supervision of contemporary clinical surveillance practice and is preferred for general surveillance because it is comprehensive and assistance-oriented, in clinical supervision Supervisors are more willing to assist and cooperate with the diagnosis and prescribing process (Sarfo and Cudjoe 2016), which was modified as a driving force for professional teachers to teach. (Okorji and Ogbo 2013). Although in practice it is still rarely carried out by supervisors because it is quite time-consuming and there are still teachers who think that clinical supervision is only for looking for teacher weaknesses (Veloo et al. 2013), this clinical supervision can be used as an alternative to improving teacher performance.

The results of analysis from several literature on Clinical Supervision show that there is an effect of clinical supervision on teacher teaching performance. Other findings suggest that supervisors have not provided enough clinical supervision, because most of the time they focus on administrative aspects this is one of the obstacles to implementing clinical supervision in schools. The advantages of clinical supervision over academic supervision are in the process of how to improve learning, starting from before observation, during classroom observations and the follow-up process, whereas academic supervision is sufficient to only see class observations writing down the results and rarely follow up on improvement (Bessong and Ojong 2010; Chidi and Victor 2017; Fteiha and Abdawi 2017; Gursoy et al. 2013; Okorji and Ogbo 2013; Sarfo and Cudjoe 2016; Veloo et al. 2013)

The role of the instrument is very important in supervision as a means of measuring the expected performance of actual performance, because the more valid the instrument used, the more precise the supervisory data will be collected. The function of the instrument as a performance measurement tool is to help supervisors know the effectiveness of the learning process, to find out to what extent the level of success of students in achieving learning objectives after they take the learning process. The requirement for a good measuring instrument is that it can provide accurate information to its users, which must be valid, reliable and fair, meaning that the measuring instrument can distinguish the abilities of each teacher when the assessment process occurs in learning.

Meanwhile, some of the existing instruments in the Supervisory Board still do not pay attention to the aspects of development and up to date information (Anon n.d.). This is one of the obstacles to effective and professional supervision, supervision is still only administrative tools, and rarely conducts observations of learning in the classroom (Asdhiani, Saptono, and . 2020).

The activity of supervision / supervision of Islamic Religious Education is a series of activities from the administration of Islamic Religious Education, by conducting supervision and assessment of what has been planned and implemented in Islamic Religious Education activities. The supervision process carried out by the Islamic Religious Education Supervisor does not only see the learning outcomes but how the learning process is carried out. One of the reasons is due to the development of changes in the Islamic Religious Education curriculum, so that teachers in schools need continuous adjustments to the real conditions in their respective schools.

Based on the results of preliminary studies, empirically found the following. First, some school principals, teachers and even supervisors do not fully understand what is meant by clinical supervision. Second, supervision activities are still perceived as a school principal program and a supervisory program. So that the teacher is passive in waiting when the principal or supervisor has time to supervise the teacher. Third, there is no standard instrument used by supervisors in the clinical supervision process. Fourth, in real conditions the teacher must be supervised and there are many in number, while time is limited, then through the help of practical clinical supervision instruments it can be used as a strategy to foster teachers by adopting time strategies without having to meet in real time at each stage of supervision. clinical, because by using the media of information and technology (IT) communication will continue without time constraints having to meet regularly with frequent frequency.

Errors that often occur in measurement are caused by the measuring instrument, which is being measured and who measures it. Therefore it is necessary to have a measurement theory so that it can make the right decisions according to what it is measuring, using either classical or modern theory. This study will use the Rasch modeling initiated by Dr. George Rasch, the advantage of this model is that it is able to predict the missing data based on a systematic response pattern, is able to produce a standard error measurement value for the instrument used so as to increase the accuracy of calculations, can test respondents and items simultaneously, can calibrate at once in three things (measurement scale, respondent and item) (Sumintono and Widhiarso 2015).

One of the measurement tools or assessments used to assess a person's ability is performance assessment (Mardapi 2008). Performance appraisal is a type of evaluation that requires expert feedback and ratings (Toffoli, de Andrade, and Bornia 2016). When involved in the appraisal rating action, the assessor makes observations and interpretations so that it triggers the variability of the assessor and the variability of the rating scale if there is more than one assessor, this can affect the fairness of the performance appraisal (Eckes 2019).

This research will explain the use of MFRM, because MFRM is a logistical latent trait model of probability that independently calibrates the test item difficulty and participant ability, but places them in the same frame of reference, which allows the research to add aspects of jury severity to people's abilities and item difficulties and placing them on the same logit scale (log odds unit) for comparison, the MFRM adjusts for the variability of the rater and thus provides a more accurate picture of ability (Farrokhi, Esfandiari, and Schaefer 2012).

The clinical supervision instruments owned by the Islamic Religious Education Supervisor according to the results of the observations did not have a standard instrument according to the stages in clinical supervision, namely the pre-observation conference stage, the implementation of the observation and the post-observation conference. The instrument that already exists and is implemented is the learning observation stage, for the other stages do not have the instruments. Based on the results of interviews with several Supervisors of Islamic Religious Education in Bekasi District, Bekasi City, Karawang Regency, Kuningan Regency, Bandung City and Bandung Regency, the implementation of clinical supervision is still rarely carried out in schools, due to several reasons including there are no standard instruments, if you have already the pre-observation conference stage has not been carried out directly to the observation and follow-up stage.

Literature Review

Clinic Supervision

This idea of clinical supervision was pioneered by Morris Moto in the 1960s in the writing of Case Studies and Research in Clinical Monitoring, and further enhanced by Goldhammer, Anderson & Krajewski in the 1980s (Taib et al. 2015). In the late 1970s and early 1980s, many of the major theorists in instructional supervision focused on measuring teaching behavior, accountability systems, and evaluation of teaching performance (Barbara et al. 1996). This development is due to the positivistic paradigm that dominates the field of social science research and evaluation. During the 1980s, however, American education reform introduced measurable results, prompting even more steps toward supervisory teaching practices focused on evaluation and accountability.

Supervision in the context of the school does not mean the process of supervising and improving teaching but to improve the process of learning outcomes. Supervision can be defined as activities related to teaching behavior, curriculum, learning environment, student grouping, teacher utilization, professional development, teaching improvement, and improving classroom practice for the benefit of students (Wanzare and da Costa 2000). Supervision is a service activity for teachers and students both as individuals and in groups to improve learning by providing best practices in the teaching-learning process, to improve student academic achievement, so that the feedback can help teachers apply modern methods of teaching, innovation and technology in the room. their classes, enhance their job performance, professional growth and their career development. Supervision is carried out to increase the effectiveness of learning (Anon 2013).

The best clinical supervision requires a collaborative process of teacher and observer for enhanced learning. The concept of clinical supervision according to Anderson (1986) is:

systematic investigation, developing a learning process through modified teacher behavior, planned monitoring objectives, objective data, Supervisors determine methods for class data collection to be free from bias, pattern analysis, data analyzed and organized by supervisors to describe behavior patterns that have been discussed in pre-observation and relate to teacher behavior, flexible methodologies, delineation of roles between supervisors and teachers according to performance, training clinical skills of supervisors, supervisors need to be trained not only clinical supervision skills but learning theory, teaching methodology , effective teaching, communication skills, and organizational change, productive support in a supervisory climate, because the clinical surveillance cycle implies a long-term commitment to improved teaching.

Instrument Validity

Cronbach (1971) describes validity as the process by which a test developer or test user collects evidence to support the type of conclusions to be drawn from a test score, to plan a validation study, the desired conclusions must be clearly identified, then empirical studies are designed to gather evidence of the usefulness of the score. for that (Crocker and Algina 1986), validity relates to the extent to which the results meet their intended use (John 2015)Validity is very important in evaluation, validity is the process of interpreting the test results not the test itself, concluded from existing theoretical evidence (Mardapi 2008; Sumintono and Widhiarso 2015; Wainer and Braun 2013), specifically used for specific purposes such as selection tests, evaluation of learning and others, and expressed by degrees such as high, medium or low (Norman Edward Gronlund 1982).

The validity in the Rasch model is in accordance with the model according to Hambleton and Swaminathan in Safari (2018). Analysis with the Rasch model produces a statistical analysis of

Rasch Model

Rasch modeling is an alternative to developing measurement instruments besides using classical theory. The Rasch measurement model developed not only for dichotomous data but for polytomic data as well as developed by David Andrich, then there is John Linacre who developed a multi-rater data analysis which assesses a particular instrument by more than one rater to assess the consistency of the assessors, a multi-facet model is developed. Rasch.

The Rasch Model application for instrument testing according to Linacre (Sumintono and Widhiarso 2015) is a measuring construct map, a construct map is basically a visual representation to be able to know the exact location of the items and the respondent in terms of the dimensions measured

suitability (fit statistics) which provides information on whether the data obtained ideally illustrates that people who have high ability provide patterns of answers to items according to their level of difficulty. The parameters used are the infit and outfit of the middle square (mean square) and standardized values. Infit (inlier sensitive or information weighted fit) is the sensitivity of the response pattern to the target item on the respondent (person) or vice versa; while outfit (outlier sensitive fit) measures the sensitivity of response patterns to items with a certain difficulty level for respondents or vice versa (Sumintono and Widhiarso 2015)

Instrumen Reliability

Measurement in education cannot be directly carried out on the measured character, because it is abstract, but it can be measured through an indicator. In measurements that contain psychological aspects, it is not easy to obtain accurate data, because there are several sources of error, namely measuring instruments, which are measured and those that measure. There are three things that describe the reliability of both test and non-test instrument measurements, namely, stability, equivalence, and internal consistency (Sumintono and Widhiarso 2015). This stability can be said to be reliability, to see the reliability of a measuring instrument can be done by statistical calculations, its value is called the reliability coefficient, which is the coefficient of consistency or the stability of the measurement results (Retnawati and Nugraha 2016)

Research is considered reliable if it provides consistent results for the same measurement and is considered unreliable if repeated measurements give different results. The level of reliability is empirically indicated by a number called the reliability coefficient value. It is denoted by the letter "r," and is expressed as a number that ranges between 0 and 1.00, with r = 0 indicating no reliability, and r = 1.00 indicating perfect reliability. Don't expect to find a test with perfect reliability. In general, to view test reliability as a decimal, for example, r = .80 or r = .93. The larger the reliability coefficient, the more repeatable or reliable the test score will be. However, do not select or reject tests based solely on the measure of their reliability coefficient. To evaluate the reliability of the test, one must consider the type of test, the type of reliability estimate reported, and the context in which the test will be used (Kimberlin and Winterstein 2008)

The reliability value in the Rasch model is indicated by the value of individual separation (person separation), which is how well the items in the test spread across the range, the greater the better the test is arranged because it reaches all individual abilities and item separation is how big the sample is used. scattered along a linear interval scale, the higher the value the better the measurement (Sumintono and Widhiarso 2015)

Methods

Research Model

This article uses a research & development model with the aim of producing a product in the form of an instrument for evaluating the performance of religious teachers through a clinical supervision model. The development model used in this research is the R & D development model of Borg and Gall's theory (Gall, M. D., Gall, J. P., & Borg 2003). The procedures in this development include: (1) Introduction (Define); (2) Planning (3) Developing the initial product (develop), (4) Initial testing; (5) The first revision; (6) main field trials; (7) second product revision; (8) Operational product testing; (9) Final Product Revision; (10) Presentation of the final product (Deliver).

Based on the ten steps of research and development developed by Borg & Gall, in this study, the implementation process makes an adaptation referring to the model approach. As for the adaptation of this development research in general, it consists of three main activity stages, namely: (1) Define; (2) develop; (3) deliver of product development results.

Procedure

The define stage consists of three main activities which include needs analysis, designing grids and making performance assessment instruments for religious teachers with clinical supervision models**.**

The needs assessment activity aims to reveal the real conditions of Islamic education teachers in assessing student performance, especially in the current Islamic education learning. Needs analysis is in the form of survey results using a set of questions conducted on Islamic education teachers in SD, SMP and SMA. As research subjects at the needs analysis stage, information was obtained that the constraints on the performance assessment of Islamic Education learning activities, one of the obstacles was that the teacher still did not understand the scoring guidelines in instruments that were unclear so that it was difficult to use, the components that were considered difficult to observe, so they tended to be ignored.

Develop stage, at this stage the instrument that has been designed is consulted with the supervisor. Expert test or validation, conducted with respondents who are experts in designing a model or product. This activity is carried out to review the initial product and provide input for improvement.

This validation process is called Expert Judgment. Instruments that have been produced are evaluated, whether the resulting format is appropriate or not, and how is the suitability of the content of the learning assessment material and is analyzed using CVR and CVI. CVR and CVI were proposed by Lawshe in 1975 using 3 rating scales (Lawshe, 1975)

If the instrument is not yet feasible, then the instrument is revised again so that the instrument becomes feasible to be tested. Before the trial, 21 experts validated the instrument, then the instrument was tried out to assess the teacher's performance during Islamic Education learning on a small scale totaling 20 teachers. This aims to determine whether the instrument is suitable for use or not to find out how the teacher is performing. The results of the small-scale trial using the instrument are used as a reference for further development and improvement of the instrument.

Deliver stage, at this stage it was tested more widely, namely by the number of more teachers, totaling 50 teachers. The actual product trial was carried out to assess the performance of Islamic Education teachers during the lesson. the result of this phase is the conclusion of the success or failure of the product design developed for the benefit of the user and the team involved.

Participants

The actual product trial subjects were Islamic religious education teachers at the elementary school to senior high school levels in Bekasi district, totaling 20 teachers on a small scale. Meanwhile, the large-scale trial involved 50 different Islamic education teachers from the elementary school to senior high school.

Instruments

The instruments used in this study are as follows; 1) Rubric, Rubric is used as an assessment guide that describes the criteria desired by the rater in assessing or grading the teacher's work. The rubric lists the desired characteristics that need to be demonstrated in a teacher's job accompanied by a guide for evaluating each of these characteristics. The purpose of the rubric assessment is that the supervisor (rater) is expected to clearly understand the basis for the assessment that will be used to measure teacher performance. Both parties (supervisors and teachers) will have clear shared guidelines on the expected performance demands. 2) Rating Scale, This multilevel scale contains teacher activities in the form of skills to be observed. This graded scale is filled by the rater who observes all teacher activities. 3) Observation sheet, Observation guidelines in the performance appraisal of religious teachers in the clinical supervision model are used by supervisors to observe and assess each teacher by using a rubric in the form of a rating scale and the weight of the assessment. Compilation of the construct of the observation instrument in the form of a rating scale based on material that reflects the skills to be measured. Furthermore, the rating scale for each material is determined. In this study, four scales were used, namely one to four (1-4).

Data Analysis

A range of statistical techniques such as factor analysis, Cronbach's alpha calculation, point biserial correlation, and computation of the raw total score are commonly used to develop instruments (tests, surveys) for educational research. This approach has been used to evaluate the strength of the conclusions drawn from the instrument and to calculate the performance of respondents (eg, students, teachers). Rasch analysis is a psychometric technique developed to increase the accuracy of researchers in arranging instruments, monitoring the quality of the instruments, and calculating the performance of respondents. Rasch analysis allows researchers to construct alternative forms of measurement instruments, which open the door to changing instruments in the context of student growth and change. Rasch analysis also helps researchers think in more sophisticated ways with respect to the constructs (variables) they want to measure. Some life science education researchers are already using the Rasch technique (Reeves and Marbach-Ad 2016), but many continue to use instrument development and validation approaches that rely on classical test theory.

The analysis in this study is the validity and reliability of clinical supervision instruments. Clinical Supervision assessment instruments for religious teachers will be submitted to experts to be tested for the feasibility or relevance of their contents, the results of expert assessments are carried out by analysis using CVR and CVI. CVR and CVI were proposed by the proof of content validity by Lawshe (Lawshe, 1975).

Reliability studies involving raters are usually called inter-rater agreements or inter-rater reliability. If in the case of self-report reliability is shown by internal consistency that is seen from one item to another which has a high correlation, then in the case of inter-rater reliability, the consistency is tested. So the grain position is replaced by the person position (rater). The rater who has high agreement is seen from the position of the observed subject. If the order of subject scores from Rater A and B is almost the same, then the two raters have high agreement (Ebel & Frisbie, 1991). This is because the agreement is operationalized in the form of a correlation. The rater or panelists who will be used in the assessment process are three mathematics teachers who already have certification so that later in the assessment process they can minimize the level of subjectivity.

This article uses the approach to see the reliability of the instrument using the Multifaceted Rasch Model. The Rasch model uses a probabilistic response distribution as a logistical function of the person and item parameters to determine unidimensional latent traits. In contrast to the ICC where the raw score is directly used in the analysis, the Rasch measurement theory converts the raw score to a log-odds scale using a logistic transformation. The transformed test score data can then be conceptualized as a dependent variable with several independent variables (eg Aspects) of interest, including measures of severity and leniency, item difficulty, task difficulty, and level of performance achievement. The calculation uses the FACETS software, the FACETS software is an extension used in the Rasch measurement model (Wolins, Wright, and Masters 1983; Wolins, Wright, and Rasch 1982; Wright and Stone 1979) which can be used to write assessments that cover various aspects, such as assessors and writing assignments.

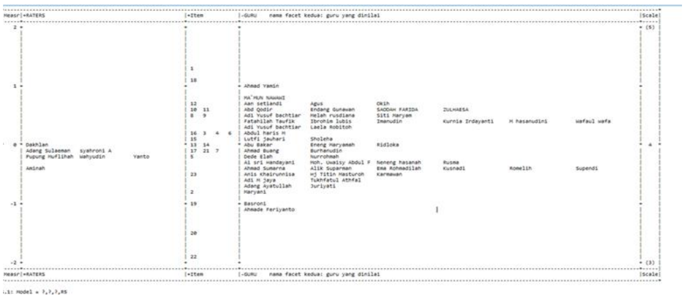
Results

The results in content validation for the clinical supervision assessment tool were analyzed using the content validity Lawshe where the CVR validity standard was 20 people. The CVR value must meet 0.5 so that the items can be declared valid. This applies to content validation using 20 SMEs (LAWSHE 1975). The CVR value obtained from each item is 1 and is fully presented in the attachment. The CVI value obtained from the average CVR is 1. Based on the CVR value that exceeds 0.5, all items are declared valid (LAWSHE 1975)and are suitable for use for further research

The implementation stages in the trial are as follows; 1.Before the trial is carried out, the researcher provides an instrument to the supervisor / rater to explain the intentions contained in the indicator items. Each supervisor / rater gets 50 instruments to fill in with a value of 1,2,3,4 and 5 so that when the supervisor / rater conducts the assessment there is no misperception or interpretation of the assessment items, 2.The supervisor / rater conducts an assessment of the Religious Education teacher Islam, 3. the researcher held a discussion with the supervisor / rater and asked for input on the assessment instrument used. The average value at the time of assessment can be seen in the appendix. The clinical supervision instrument consists of 23 statement items and in it there are 5 performance standards that have been tested for the validity of the content and the reliability. The results of the observations from each supervisor / rater were analyzed using Facets software.

The summary report produced by Facets shows that the mean and standard deviation of the standard residuals for this Rasch model are 0 and 1.Linacre (2011) emphasizes that "... when the data fits the Rasch model, the mean of the standard residuals is expected to approach 0, 0, and the sample standard deviation is expected to be close to 1.0. Therefore, this Rasch model is useful for explaining rater severity. The results of data analysis mean 3.96 means that the ability of the respondents is high towards the item.

The figure 1 above shows the distribution of three aspects in this study, namely; the severity of the rater (Rater), namely the Islamic Religious Education supervisor of seven people, the instrument items totaling 23 indicators and the Islamic Religious Education teacher (ratee) totaling 50 people. The logit scale is shown in the far left column where average ability / difficulty is set to zero. Negative scores indicate low proficiency for the test taker, leniency for the rater, and facilities for the item, while positive scores indicate high proficiency, severity, and difficulty. The far right column displays the corresponding raw score scale.



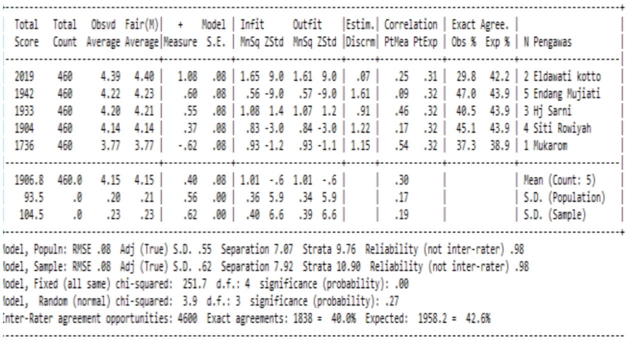
```
|Measr|+RATERS                           |+Item      |-GURU    nama facet kedua: guru yang dinilai                                                    |Scale|
+   2 +                                  +           +                                                                                          + (5) |
|     |                                  | 1         |                                                                                          |     |
|     |                                  | 18        |                                                                                          |     |
+   1 +                                  +           + Ahmad Yamin                                                                              +     +
|     |                                  |           | MA'MUN NAWAWI                                                                            |     |
|     |                                  | 12        | Aan setiandi        Agus                Okih                                             |     |
|     |                                  | 10  11    | Abd Qodir           Endang Gunawan      SAODAH FARIDA      ZULHAESA                      |     |
|     |                                  | 8   9     | Adi Yusuf bachtiar  Melah rusdiana      Siti Maryam                                      |     |
|     |                                  |           | Fatahilah Taufik    Ibrohim lubis       Imanudin           Kurnia Irdayanti   M hasanudini   Wafaul wafa |     |
|     |                                  |           | Adi Yusuf bachtiar  Laela Robitoh                                                        |     |
|     |                                  | 16  3  4  6 | Abdul haris M                                                                          |     |
|     |                                  | 15        | Lutfi jauhari       Sholeha                                                              |     |
*   0 * Dakhlan                          * 13  14    * Abu Bakar           Eneng Maryamah      Ridloka                                      *  4  *
|     | Adang Sulaeman   syahroni A      | 17  21  7 | Ahmad Buang         Burhanudin                                                           |     |
|     | Pupung Muflihah  Wahyudin   Yanto| 5         | Dede Elah           Nurrohmah                                                            |     |
|     |                                  |           | Ai sri Handayani    Moh. Uwaisy Abdul F Neneng hasanah     Rusma                         |     |
|     | Aminah                           |           | Ahmad Sumarna       Alik Suparman       Ema Rohmadilah     Kusnadi     Romelih   Supendi |     |
|     |                                  | 23        | Anis Kheirunnisa    Hj Titin Masturoh   Karmawan                                         |     |
|     |                                  |           | Adi M Jaya          Tukhfatul Athfal                                                     |     |
|     |                                  |           | Adang Ayatullah     Juriyati                                                             |     |
|     |                                  | 2         | Maryani                                                                                  |     |
+  -1 +                                  + 19        + Basroni                                       |                          +     +
|     |                                  |           | Ahmade Feriyanto                                                                         |     |
|     |                                  | 20        |                                                                                          |     |
|     |                                  | 22        |                                                                                          |     |
+  -2 +                                  +           +                                                                                          + (3) +
|Measr|+RATERS                           |+Item      |-GURU    nama facet kedua: guru yang dinilai                                                    |Scale|
.1: Model = ?,?,?,RS
```

The useful information in the image above is the Appraiser according to the picture above, namely the supervisors of Islamic Religious Education, Eldawati Koto and Endang Mujiawati including severe / savere assessors / supervisors who provide a value, while Mukarom is an assessor / supervisor of Islamic Religious Education who is loose / lenient giving a value , while Hj Rowiyah and Hj Sarni were among the supervisors who were moderate in giving scores.

Measure was changed to 2, 1, 0, -1 and -2, meaning that the abilities of Islamic Religious Education teachers were normally distributed because most were in the logit range of -1 to 1, Islamic Religious Education teachers named Rasmin and Styaka got the highest scores according to the Supervisors. , Encin and Dedeh got the lowest score from the supervisors, the item indicator of the instrument that was most difficult to answer by Islamic Religious Education teachers was number 6 but was able to be answered by Styaka and Rasmin and the easiest was number 2. In the classical analysis did not explain in detail about participants, appraisers or items, whereas in modern analysis through multi facets, according to Bambang Sumintono, it is the opposite.

Supervisor number 3 Dakhlan is an assessor who is not in the severe category because the measure value is 0.00 log, rater / supervisor number 1 Aminah is a loose assessor with measure-0.41 log. The criteria for fit are Outfit means square (0.5 <MNSQ <1.5), Outfit Z-standard (-2 <ZTSD <+2), and Point measure correlation (0.4 <Pt Measure Corr <0.85). . In the picture above all supervisors meet the fit criteria, for ZTSD there is one supervisor who meets the misfit criteria, namely Aminah 2.7 log and 2.5 log, there are two supervisors who do not meet the point measure correlation Pupung Muflihah 0.3 and Aminah 0.26 as note that the ZTSD value is very much affected by the sample size. When the sample size is large it is ensured that the ZTSD value is always above 3, according to experts recommend not using

ZTSD if the sample size is> 500 (Sumintono and Widhiarso 2015)



```
+---------------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair(M)|   +     Model | Infit      Outfit    |Estim.| Correlation | Exact Agree. |                  |
| Score   Count  Average Average|Measure  S.E.  | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Obs %  Exp % | N Pengawas       |
|-------------------------------+---------------+----------------------+------+-------------+--------------+------------------|
| 2019     460     4.39   4.40  |  1.08    .08  | 1.65  9.0  1.61  9.0 |  .07 |   .25   .31 |  29.8   42.2 | 2 Eldawati kotto |
| 1942     460     4.22   4.23  |   .60    .08  |  .56 -9.0   .57 -9.0 | 1.61 |   .09   .32 |  47.0   43.9 | 5 Endang Mujiati |
| 1933     460     4.20   4.21  |   .55    .08  | 1.08  1.4  1.07  1.2 |  .91 |   .46   .32 |  40.5   43.9 | 3 Hj Sarni       |
| 1904     460     4.14   4.14  |   .37    .08  |  .83 -3.0   .84 -3.0 | 1.22 |   .17   .32 |  45.1   43.9 | 4 Siti Rowiyah   |
| 1736     460     3.77   3.77  |  -.62    .08  |  .93 -1.2   .93 -1.1 | 1.15 |   .54   .32 |  37.3   38.9 | 1 Mukarom        |
|-------------------------------+---------------+----------------------+------+-------------+--------------+------------------|
| 1906.8  460.0    4.15   4.15  |   .40    .08  | 1.01  -.6  1.01  -.6 |      |   .30       |              | Mean (Count: 5)  |
|   93.5     .0     .20    .21  |   .56    .00  |  .36  5.9   .34  5.9 |      |   .17       |              | S.D. (Population)|
|  104.5     .0     .23    .23  |   .62    .00  |  .40  6.6   .39  6.6 |      |   .19       |              | S.D. (Sample)    |
+---------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .08  Adj (True) S.D. .55  Separation 7.07  Strata 9.76  Reliability (not inter-rater) .98
Model, Sample: RMSE .08  Adj (True) S.D. .62  Separation 7.92  Strata 10.90  Reliability (not inter-rater) .98
Model, Fixed (all same) chi-squared:  251.7  d.f.: 4  significance (probability): .00
Model,  Random (normal) chi-squared:  3.9  d.f.: 3  significance (probability): .27
Inter-Rater agreement opportunities: 4600  Exact agreements: 1838 =  40.0%  Expected:  1958.2 =  42.6%
-----------------------------------------------------------------------------------------------------------
```

Figure 1. Measurement Report

Figure 2 below shows the rater severity rating output provided by the FACETS program. All three raters had an acceptable fit according to the Disability and Achievement measures well in the range 1.50 to 0.50, indicating good consistency in the ratings. In the context of rater terms, the fit index can be interpreted as a measure of intra-rater reliability, where the difference between the fit measure and the optimal value of 1.00 indicates the percentage of unexplained noise in the response pattern (Wright & Linacre, 1994).

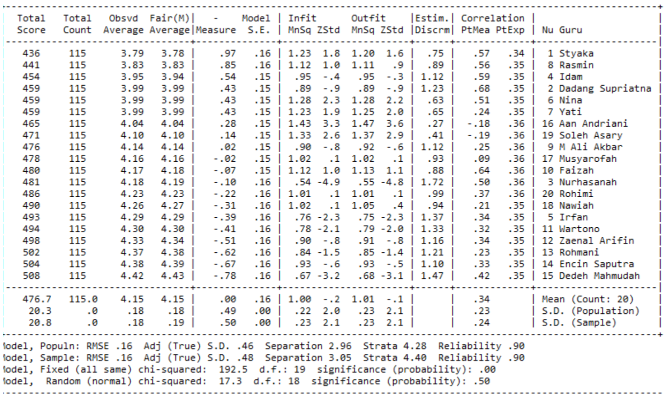
Apart from the outliers shown in the variable map, the personification measure showing all of the test participant's Fit and Outfit indices falls within the acceptable range of 0.50-1.50 (Linacre and Wright 1994), which means MFRM can reliably estimate ability for nearly any individual. . A split index of 3.21 and a reliability score of 0.98 suggest this test may be useful when applied to other test takers from the same population (Wright and Stone 1979). In particular, the test was able to assess broad variability in the ability levels of test takers, differentiating between the five supervisor ability groups.

The Fixed effect X2 indicates that the people tested are statistically different, that is, independence in the parameters of the person applies (X2 = 251.7, df = 4, p <0.000). In addition, the comparison of the Fair-M estimate (M = 4.15, SD = 0.21) with the observed mean (M = 4.15, SD = 0.20) further confirms the lack of different scores by the experts. That is, the mean raw score requires minimal adjustment to compensate for the effects of rater variability.

Furthermore, the separation index of 7.92 and reliability of 0.98 indicate that this rater indicates sufficient inter-rater reliability. The FACETS program calculates reliability as a measure of variance in the sample; therefore, low values among the rater's sample mean relatively homogeneous raters in the assessment, a desirable feature from a rater's perspective and an indicator of convergent validity. Although the separation index of 7.92 can be considered low, being greater than 1.00 indicates that the three raters are close to heterogeneous rater

Figure 3 shows all the Fit and Outfit indexes of Islamic Religious Education teachers are in the acceptable range of 0.50-1.50 (Wright & Linacre, 1994), which means that the Multi Facets Rasch Model can reliably estimate the ability for almost all individual teachers. The mean value of 0.00 and the SD value of 0.49 means that the measure rate / teacher value is above 0.49 including the Islamic Religious Education teacher whose learning performance is high, namely Sytaka from Junior high school, Rasmin. Idam, measure values 0.49 to -0.49 ratee have moderate learning performance, namely Dadang, Nina, Yati, Aan, Soleh, M. Ali, Musyarofah, Faizah, Nurhasanah, Rohimi, Nawiah, Irfan, Wartono, while the measure values are below - 0.59 namely Zaenal, Rohmani, Encih and Dedeh including teachers who have low learning performance.

The infit and outflow values of MnSq are between 0.5 to 1.5 log as many as 20 teachers, which means that they show good consistency in assessment. A reliability score of 0.90 means good. The separation index of 2.96 can be considered low because it is greater than 1.00, it indicates that the twenty teachers have a fairly heterogeneous performance in learning performance.



```
+-------------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair(M)|    -    Model | Infit       Outfit     |Estim.| Correlation |          |
| Score   Count  Average Average|Measure  S.E.  | MnSq ZStd   MnSq ZStd  |Discrm| PtMea PtExp | Nu Guru  |
|-------------------------------+---------------+------------------------+------+-------------+----------|
|  436     115      3.79   3.78 |    .97   .16  | 1.23  1.8   1.20  1.6  |  .75 |    .57   .34 |  1 Styaka |
|  441     115      3.83   3.83 |    .85   .16  | 1.12  1.0   1.11   .9  |  .89 |    .56   .35 |  8 Rasmin |
|  454     115      3.95   3.94 |    .54   .15  |  .95  -.4    .95  -.3  | 1.12 |    .59   .35 |  4 Idam   |
|  459     115      3.99   3.99 |    .43   .15  |  .89  -.9    .89  -.9  | 1.23 |    .68   .35 |  2 Dadang Supriatna |
|  459     115      3.99   3.99 |    .43   .15  | 1.28  2.3   1.28  2.2  |  .63 |    .51   .35 |  6 Nina   |
|  459     115      3.99   3.99 |    .43   .15  | 1.23  1.9   1.25  2.0  |  .65 |    .24   .35 |  7 Yati   |
|  465     115      4.04   4.04 |    .28   .15  | 1.43  3.3   1.47  3.6  |  .27 |   -.18   .36 | 16 Aan Andriani |
|  471     115      4.10   4.10 |    .14   .15  | 1.33  2.6   1.37  2.9  |  .41 |   -.19   .36 | 19 Soleh Asary |
|  476     115      4.14   4.14 |    .02   .15  |  .90  -.8    .92  -.6  | 1.12 |    .25   .36 |  9 M Ali Akbar |
|  478     115      4.16   4.16 |   -.02   .15  | 1.02   .1   1.02   .1  |  .93 |    .09   .36 | 17 Musyarofah |
|  480     115      4.17   4.18 |   -.07   .15  | 1.12  1.0   1.13  1.1  |  .88 |    .64   .36 | 10 Faizah |
|  481     115      4.18   4.19 |   -.10   .16  |  .54 -4.9    .55 -4.8  | 1.72 |    .50   .36 |  3 Nurhasanah |
|  486     115      4.23   4.23 |   -.22   .16  | 1.01   .1   1.01   .1  |  .99 |    .37   .36 | 20 Rohimi |
|  490     115      4.26   4.27 |   -.31   .16  | 1.02   .1   1.05   .4  |  .94 |    .21   .35 | 18 Nawiah |
|  493     115      4.29   4.29 |   -.39   .16  |  .76 -2.3    .75 -2.3  | 1.37 |    .34   .35 |  5 Irfan  |
|  494     115      4.30   4.30 |   -.41   .16  |  .78 -2.1    .79 -2.0  | 1.33 |    .32   .35 | 11 Wartono |
|  498     115      4.33   4.34 |   -.51   .16  |  .90  -.8    .91  -.8  | 1.16 |    .34   .35 | 12 Zaenal Arifin |
|  502     115      4.37   4.38 |   -.62   .16  |  .84 -1.5    .85 -1.4  | 1.21 |    .23   .35 | 13 Rohmani |
|  504     115      4.38   4.39 |   -.67   .16  |  .93  -.6    .93  -.5  | 1.10 |    .33   .35 | 14 Encin Saputra |
|  508     115      4.42   4.43 |   -.78   .16  |  .67 -3.2    .68 -3.1  | 1.47 |    .42   .35 | 15 Dedeh Mahmudah |
|-------------------------------+---------------+------------------------+------+-------------+----------|
|  476.7   115.0    4.15   4.15 |    .00   .16  | 1.00  -.2   1.01  -.1  |      |    .34      | Mean (Count: 20) |
|   20.3      .0     .18    .18 |    .49   .00  |  .22  2.0    .23  2.1  |      |    .23      | S.D. (Population) |
|   20.8      .0     .18    .19 |    .50   .00  |  .23  2.1    .23  2.1  |      |    .24      | S.D. (Sample) |
+-------------------------------------------------------------------------------------------------------+
odel, Populn: RMSE .16  Adj (True) S.D. .46  Separation 2.96  Strata 4.28  Reliability .90
odel, Sample: RMSE .16  Adj (True) S.D. .48  Separation 3.05  Strata 4.40  Reliability .90
odel, Fixed (all same) chi-squared:  192.5  d.f.: 19  significance (probability): .00
odel,  Random (normal) chi-squared:  17.3  d.f.: 18  significance (probability): .50
---------------------------------------------------------------------------------------------------------
```

Figure 3. Fit and Outfit indexes of Islamic Religious Education teachers

Discussions

From this study we conclude that the Rasch measure provides a detailed analysis of several variables that have the potential to impact the test or assessment results. The Rasch model has various advantages over the psychometric approach in the item response theory (IRT) framework, the most important of which is the invariance of measurement or specific objectivity. The invariance of the measurement implies the following: (a) the test score is a statistic sufficient to estimate the size of the examinee i.e. the number of correct scores of the examinee contains all the information necessary to estimate the examinee's size from a given set of observations, and (b) the test is unidimensional, that is, all items on the test measure the same latent variables or constructs (Eckes 2019). The MFRM model meets the objectivity requirements that are equivalent to other Rasch models, because each facet parameter is estimated independently from other facets so that the ability of the examinee does not depend on the item and the rater (Toffoli et al. 2016). The MFRM allowed the study to add aspects of jury severity to people's abilities and item difficulties and to place them on the same logit scale (log odds unit) for comparison. (Farrokhi et al. 2012).

Research on the development of this clinical supervision assessment instrument The Islamic Religious Education Supervisor assessed the quality of responses built based on their understanding of the performance / performance of Islamic Religious Education teachers in learning by utilizing an assessment rubric. In the assessment process there may be rater variability, rater variability is a component of undesirable variability contributing to irrelevant construct variants in the test taker's score, variability can obscure the construct being measured and can threaten the validity and fairness of performance appraisal other terms namely the rater effect, rater error or rater bias (Eckes 2019).

Cronbach (Chidi and Victor 2017) argues that the most serious error effect that can be caused by the rater is the leniency / severity effect (Myford and Wolfe 2004). The results of large-scale research analysis data assessors / supervisors of Islamic Education which severely have a value of 0.00, namely Dakhlan, which means the assessor / supervisor of Islamic Religious Education has no severity value and the loose one has a value of -0.41, namely Aminah which means the assessor / supervisor of Islamic Religious Education. has a looseness value, a separation value of 1.57 and a reliability of 0.71 means that it is moderate.

Item quality was evaluated on two aspects, namely the overall quality as indicated by the logit score and its suitability statistic (Zhu, Ennis, and Chen 1998). A logit on the score will show the overall rating of the raters on an item in terms of representativeness of its content. The item suitability statistics will reflect the unexpected ranking level of an item, if there is a severe appraiser giving a high rating on an item and many soft appraisers give an item a low rating it will be defined as a mismatched item or a misfit item, an item must be revised or deleted. if they have low logit scores or poor suitability statistics, all low and inconsistent items will be identified by Rasch modeling in the MFRM model (Fox and Jones 1998).

Ultimately, we would like to show that this study is a first step to explore the possible effects on the development of a clinical supervision assessment instrument in Islamic Religious Education teachers from the effects of severe, moderate and lenient assessment, validating and reliable clinical supervision assessment instruments. the Islamic Religious Education teacher. This research has implications, one of which is the suggestion of training for assessors / supervisors of Islamic Religious Education in the assessment because supervisors in assessing teacher performance very often use the instrument, although the training of assessors cannot eliminate the mistakes made by the assessors in terms of subjectivity, but it can provide reinforcement of the theory. required in the assessment. This research is limited in many ways, including this research is purely quantitative adding a qualitative component to provide deeper insights into the findings obtained

Limitations and Future Studies

As this study focuses on research and practice-based models, the material has the potential to benefit other higher education institutions offering teacher education programs. Although this study mostly focuses on assessing teacher performance in several schools, the proposed scoring system can be adapted to the assessment of supervisors in other disciplines and can be replicated in other institutions. The requested assignment and the rubric always remain the same; all changes are standardized test questions and student responses. A number of regions in Indonesia should routinely release items from the open standard test and provide a graded sample of teacher responses to these items. Therefore, to implement the proposed scoring system in other institutions, assessors only need to select the standard test items released and the teacher's response to those items.

Perhaps what is needed is a more mathematically sophisticated approach to adjustment that will take into account the potentially localized nature of the rater's effect of severity. Such an approach would not make the assumption that a heavy rater maintains a constant level of severity, no matter what the rating is, no matter what day the rating occurs, no matter whether the ratinge is graded in the morning or evening. , no matter whether rater is rated first or last, no matter what subgroup rater belongs to, etc. In contrast, this alternative approach would consider these contextual (and potentially powerful) aspects of the rating operation and would use information about differences in assessors' performance in relation to these aspects in adjusting ratings. Therefore, additional research is needed to determine how best to interpret these types of interaction effects and their impact on adjustment for rater severity.

References

[1] Anon. 2013. “Skills and Attributes of Instructional Supervisors: Experience from Kenya.” *Educational Research and Reviews*.

[2] Anon. n.d. “Nur Aedi, 2014.”

[3] Asdhiani, Yuni, Ari Saptono, and . Komarudin. 2020. “Profesional Supervision Model: Development of Clinical Supervision Instruments for Teachers of Islamic Education Through a Multi-Faceted Rasch Model.” *KnE Social Sciences*.

[4] Barbara, L., G. Larry, Barbara L. White, and Larry G. Daniel. 1996. “Views of Instructional Supervision: What Do the Textbooks Say?” *Paper Presented at the Annual Meeting of the Mid-South Educational Research*.

[5] Bessong, F. .., and F. Ojong. 2010. “Supervision as an Instrument of Teaching – Learning Effectiveness: Challenge for the Nigerian Practice.” *Global Journal of Educational Research*.

[6] Chidi, Nnebedum, and Akinfolarin Akinwale Victor. 2017. “Principals’ Supervisory Techniques as Correlates of Teachers’ Job Performance in Secondary Schools in Ebonyi State, Nigeria.” *International Journal for Social Studies*.

[7] Crocker, L., and J. Algina. 1986. "What Is Test Theory ?" in *Introduction to modern and classical test theory*.

[8] Eckes, Thomas. 2019. "Many-Facet Rasch Measurement." in *Quantitative Data Analysis for Language Assessment Volume I*.

[9] Farrokhi, Farahman, Rajab Esfandiari, and Edward Schaefer. 2012. "A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in Three Types of Assessment." *JALT Journal*.

[10] Fox, Christine M., and James A. Jones. 1998. "Uses of Rasch Modeling in Counseling Psychology Research." *Journal of Counseling Psychology*.

[11] Fteiha, Ahmad, and Noor. Abdawi. 2017. "The Effectiveness of Clinical Supervision on Technology Teacher'S Professional Development in Jerusalem a Case Study." *International Conference on Research in Education and Science*.

[12] Gall, M. D., Gall, J. P., & Borg, W. R. 2003. *Educational Research: An Introduction (7th Ed.)*.

[13] Gursoy, Esim, Nermin Bulunuz, Sehnaz Baltaci Goktalay, Mizrap Bulunuz, John Kesner, and Umut Salihoglu. 2013. "Clinical Supervision Model to Improve Supervisory Skills of Cooperating Teachers and University Supervisors during Teaching Practice." *Hacettepe Universitesi Egitim Fakultesi Dergisi-Hacettepe University Journal of Education*.

[14] John, A. C. 2015. "Reliability and Validity : A Sine Qua Non for Fair Assessment of Undergraduate Technical and Vocational Education Projects in Nigerian Universities." *Journal of Education and Practice*.

[15] Kären Matison Hess, Christine Hess Orthmann. 2012. *Management and Supervision in Law Enforcement, Sixth Edition*. Sixth. Delmar.

[16] Kemendikbud. 2016. "Standar Proses Pendidikan Dasar Dan Menengah." *Lampiran Permendikbud*.

[17] Kimberlin, Carole L., and Almut G. Winterstein. 2008. "Validity and Reliability of Measurement Instruments Used in Research." *American Journal of Health-System Pharmacy*.

[18] LAWSHE, C. H. 1975. "A QUANTITATIVE APPROACH TO CONTENT VALIDITY." *Personnel Psychology*.

[19] Linacre, JM, and B. D. Wright. 1994. "Dichotomous Mean Square Chi-Square Fit Statistics." *Rasch Measurement Transactions1*.

[20] Linda Openshaw. 2012. "CHALLENGES IN CLINICAL SUPERVISION." *NACSW Convention*.

[21] Mardapi, Djemari. 2008. "Teknik Penyusunan Instrumen Tes Dan Nontes." *Yogyakarta: Mitra Cendekia*.

[22] Meriza, Iin. 2018. "Pengawasan (Controling) Dalam Institusi Pendidikan." *Jurnal Ilmiah Pendidikan Agama Islam*.

[23] Myford, Carol M., and Edward W. Wolfe. 2004. "Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II." *Journal of Applied Measurement*.

[24] Norman Edward Gronlund. 1982. "Contructing Achievement Tests." 1982.

[25] Okorji, P. N., and R. N. Ogbo. 2013. "Effects of Modified Clinical Supervision on Teacher Instructional Performance." 4(6):901–5.

[26] Olibie, Eyiuche Ifeoma, and Kate Oge Ezeoba. 2016. "Principles and Actions for E-Learning Integration in Nigerian Universities' Curriculum Delivery." *International Journal of Technologies in Learning*.

[27] Reeves, Todd D., and Gili Marbach-Ad. 2016. "Contemporary Test Validity in Theory and Practice: A Primer for Discipline-Based Education Researchers."

CBE Life Sciences Education.

[28] Retnawati, Heri, and Ariadie Chandra Nugraha. 2016. "Vocational High School Teachers ' Difficulties in Implementing the Assessment in Curriculum 2013 in Yogyakarta Province of Indonesia." *International Journal of Instruction* 9(1):33–48.

[29] Sarfo, Frederick Kwaku., and Benjamin. Cudjoe. 2016. "Supervisors ' Knowledge and Use of Clinical Supervision to Promote Teacher Performance in Basic Schools." *International Journal of Education and Research*.

[30] Sumintono, Bambang, and Wahyu Widhiarso. 2015. *Aplikasi Permodelan Rasch Pada Assessment Pendidikan*.

[31] Taib, Mohd Radzi, Zuraidah Abdullah, Nik Mustafa Mat Ail, Mohd Razi Yahya, and Norhesham Mat Jusoh. 2015. "Clinical Supervision of Teaching Mara Junior Science College (MJSC), Malaysia." *Procedia - Social and Behavioral Sciences*.

[32] Toffoli, Sonia Ferreira Lopes, Dalton Francisco de Andrade, and Antonio Cezar Bornia. 2016. "Evaluation of Open Items Using the Many-Facet Rasch Model." *Journal of Applied Statistics*.

[33] Veloo, Arsaythamby, Mary Macdalena A. Komuji, and Rozalina Khalid. 2013. "The Effects of Clinical Supervision on the Teaching Performance of Secondary School Teachers." *Procedia - Social and Behavioral Sciences*.

[34] Wainer, Howard, and Henry I. Braun. 2013. *Test Validity*.

[35] Wanzare, Zachariah, and Jose L. da Costa. 2000. "Supervision and Staff Development: Overview of the Literature." *NASSP Bulletin*.

[36] Wolins, Leroy, Benjamin D. Wright, and Geoffrey N. Masters. 1983. "Rating Scale Analysis: Rasch Measurement." *Journal of the American Statistical Association*.

[37] Wolins, Leroy, Benjamin D. Wright, and Georg Rasch. 1982. "Probabilistic Models for Some Intelligence and Attainment Tests." *Journal of the American Statistical Association*.

[38] Wright, Benjamin D., and Mark H. Stone. 1979. *Best Test Design. Rasch Measurement.*

[39] Zahraini, Dr., Benyamin Situmorang, and Tiur Asi Siburian. 2018. "Model of Education Quality Management of Traditional Islamic Boarding Schools in Aceh."

[40] Zhu, Weimo, Catherine D. Ennis, and Ang Chen. 1998. "Many-Faceted Rasch Modeling Expert Judgment in Test Development." *Measurement in Physical Education and Exercise Science*.