JTET

http://penerbit.uthm.edu.my/ojs/index.php/jtet
ISSN 2229-8932 e-ISSN 2600-7932

Journal of Technical Education and Training

Exploring Test Concept and Measurement Through Validity and Reliability Process in TVET Research: Guideline for The Novice Researcher

Kahirol Mohd Salleh[1], Nor Lisa Sulaiman[1], Gene Gloeckner[2]

[1]Faculty of Technical and Vocational Education
Universiti Tun Hussein Onn Malaysia, MALAYSIA

[2]School of Education
Colorado State University, U.S.A

*Corresponding Author

DOI: https://doi.org/10.30880/jtet.2023.15.01.022
Received 25 September 2022; Accepted 5 October 2022; Available online 31 March 2023

Abstract: Validity and reliability are important aspects of research, especially in social science and TVET research. The scales used in research that are important factors to enable good research results. The objective of this paper is to provide insight into important concepts of validity and reliability and also to introduce the major methods to assess validity and reliability as they relate to social science and TVET research. The research design and approach used for this study was a non-experimental research design. The data are using retrospective data and descriptive statistics which were used to run the simulation and analysis. The outcome of this research suggests that Cronbach's Alpha is the best method to use when checking the reliability in terms of internal consistency. The alpha provides various information that can benefit the researcher, especially in novice research, The result from the simulation shows the importance of conducting validity and reliability test during the research process. It not only provided evidence but also confidence because it reflected consistency and accuracy.

Keywords: Validity, reliability, measurement, quantitative research, TVET

1. Introduction

Research is a scientific process that involves a specific procedure in identifying, locating, assessing, and analysing all the information gathered to support and answer research problems and research questions. As the researcher progresses from one step to the next, it is often necessary to rethink, revise, and add additional material or even adjust the research perspective. The scientific process in research provides a structure that helps the researcher break down the research project into specific tasks and set deadlines, while also showing how the research is connected and built on each other (Baimyrzaeva, 2018). Much will depend on what the researcher discovers during the process. One of the most important steps is checking the validity and reliability. Validity and reliability are concepts used to evaluate the quality of research. Validity is an evolving complex concept because it relates to the inferences regarding the assessment or test results. Or in other words, is the instrument really measuring what the researcher thinks it is measuring? While reliability is dealing with research instrument consistency and replicability over time. Thus, most novice researchers struggle with this topic and struggle with differentiating between validity and reliability.

During the research design, methods planning, and results writing stages, it is important for the researcher to consider reliability and validity, especially in quantitative research. Validity is defined as the extent to which empirical evidence and theoretical concept are accurately measured in a quantitative study (Haele & Twycross, 2015). Similarly, Drost

*Corresponding author: kahirol@uthm.edu.my

(2011) defined validity as something to be concerned with the meaningfulness of research components. However, the most accurate definition of validity and reliability is from the American Educational Research Association [AERA] (2014) purported that validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity can be seen as the core of any form of assessment that is trustworthy and accurate. Validity is the extent to which the result really measures what it is supposed to measure. In the research context, any research instrument is only valid when it is measuring what is supposed to measure or when it accurately measures any prescribed variable it is considered a valid instrument for that variable. In contrast, reliability is the extent to which the result can be reproduced when the research is repeated under the same conditions. The reliability of an instrument is closely associated with its validity.

Reliability is a process of analysing the quality of the measurement process which is utilized for data collection in social science and Technical and Vocational Education and Training (TVET) research. Reliability refers to how well the items in the research instrument consistently measure what the item is supposed to measure. Reliability is referring to a measurement that supplies results that are consistent with equal values (Haradhan, 2017). Reliability is also an essential factor for data analysis. In simple terms, reliability is the degree to which research methodology produces stable and consistent results. If a measure is reliable, it means that if the measure is given multiple times, the results will be consistent each time. Reliability also addresses the consistency of an instrument from beginning to end. For example, if one was measuring "stress" with ten questions or an item questionnaire, one would expect similar results from even and odd questions.

Validity and reliability are a way of assessing the quality of the measurement procedure used in research methods and data collection in social science and TVET research. TVET is a discipline of education and training that provides capable and competent human resources with technical and vocational expertise in the job market (Salleh & Suliman, 2015). Consequently, validity and reliability become a such important concept that has been defined in terms of their application to research activities. Thus, the aim of this paper is to provide insight into these two important concepts of validity and reliability and to introduce the major methods to assess validity and reliability as they relate to social science and TVET research. The validity and reliability are abstract topics to understand especially for novice researchers. Thus, this paper provides an understanding of complex areas of validity and probability, and it has been written for novice researchers and students in the social sciences and TVET.

2. Measurement of Validity and Reliability

In general, any research instrument needs to simultaneously be valid and reliable. When choosing a research instrument or developing a new research instrument for a study, a researcher is expected to consider the relevance of the instrument to research questions as well as the quality of the instrument. This is to make sure the research instrument used is valid and reliable. An instrument's reliability is given by its consistency in measuring a fact while validity is a process of gathering information about the appropriateness of inference (Salvia & Ysseldyke, 1998).

The Standard for Education and Psychological Testing has been used as a validation criterion from 1966 to 1999. In 1999 Standards validity is measured using evidence based on content, response processes, internal structure, relations to other variables, and consequences. This includes three types of validity, Content Validity, Criterion-related Validity, and Construct validity (Gliner, Morgan & Leech, 2017). Before that, the validation standard was using the 1985 standard which used Content-related evidence, Criterion-related evidence, and Construct-related evidence. Details of the comparison between the 1999 Standards and 1985 Standards are shown in Table 1.

Table 1 - Comparison between 1999 and 1985 standards (Gliner, Morgan & Leech, 2017)

1999 Standards	1985 Standards
Evidence-based on test content	Content-related evidence
Evidence-based on response processes	Construct-related evidence
Evidence-based on the internal structure	Construct-related evidence
Evidence-based on relations to other variables	Criterion-related evidence and Construct-related evidence
Evidence-based on consequences	None

3. Measurement Concept

Before 1999, based on The Standards for Educational and Psychological Testing used the 1985 standards for validity measurement and it was traditionally subdivided into three categories including content-related, criterion-related, and construct-related validity. However, these categories change after the 1999 standards were published, validity is divided into five categories including test content, response process, internal structure, relations to other variables, and consequences. In 2014 Standards, refines the explanations of each validity with examples for each of the validity listed and integrate the validity evidence with the standards.

Table 2 - Comparison of validity and reliability in social science research (AERA, APA & NCME, 2014)

Validity / Reliability Component	Definition	Propose Statistical Analysis
Evidence-based on test content	This form of evidence is used to demonstrate that the content of the test is related to the learning that it was intended to measure.	Subject matter experts; Cohen's Kappa Index
Evidence-based on response processes	This form of evidence is used to demonstrate that the assessment requires participants to engage in specific behavior deemed necessary to complete a task.	Subject matter experts; Literature review; Content Validity Ratio; Q-sorting Scaling
Evidence-based on the internal structure	This form of evidence demonstrates how the relationships between scores on individual test items align with the construct(s) that are being measured.	Correlation; Exploratory Factor Analysis; Confirmatory Factor Analysis
Evidence-based on relations to other variables	This form of evidence demonstrates that a score measuring a defined construct relates to other scores measuring that same construct (convergent) and does not relate to other scores measuring different constructs (divergent).	Multiple Indicator-Multiple Causes Models; Principal Component Analysis; Confirmatory Factor Analysis; Q-sorting Scaling
Evidence-based on consequences	This form of evidence describes the extent to which the consequences of the use of the score are congruent with the proposed uses of the assessment.	Regression Analysis, Discriminant Analysis
Reliability	Test-Retest – to measure the consistency of results when repeating the same test on the same sample at a different point in time and to estimate the stability of test scores over time.	Intraclass Correlation Coefficient (ICC)
	Parallel forms / Alternate forms – to measure the correlation between two equivalent test versions and estimate the consistency scores across test forms.	Split-Half Coefficient; Spearmen-Brown Coefficient
	Internal consistency - the extent to which a measurement of a phenomenon provides a stable and consistent result.	Cronbach's alpha; Inter-item correlations.
	Interrater – to measure the degree of agreement between different raters/judges/assessor observing or assessing the same thing.	Cohen's Kappa; Intraclass Correlation Coefficient (ICC)

3.1 Content Validity

Content-related validity is also another type of validity. As its name implies it explores how the content of the assessment performs. Content validity is defined which the test's items represent the domain of the construct to be measured (Salvia & Ysseldyke, 1998). Similarly, content validity can also be defined as the degree to which items in the research instrument reflect the content universe to which the research instrument will be generalized (Taherdoost, 2016). Content validity includes any validity strategies that focus on the content of the test. These four elements of content validity include construct definition, construct representation, construct relevance, and appropriateness of test construction procedures (Sireci & Faulkner-Bond, 2013). In social sciences or educational research, content validity can provide detailed descriptions of the content areas, sub-content areas, and content standards. This can be done using experts' assessment or evaluation of the content of the measure, including items, tasks, formats, wording, and other criteria. Content validity is a process in which the content of a measure represents a specified content domain (Goodwin & Leech, 2003). The question can be asked, "Is the test fully representative of what it aims to measure?" An example of content validity is test questions. Does the test question really measure the item the instructor thinks the question is measuring?

3.2 Response Process Validity

Another type of validity is responses process validity which covers part of the content or construct validity. Construct, dimension, domain or latent variables can be defined as labels used to describe an unobserved behaviour that cannot be directly measured or represent something that is abstract such as feeling. The instrument must be relevant, appropriate, and utilized correctly, with the focal point being the integration of evidence. Construct validity is concerned with the efficacy of a test to gauge learner knowledge about the relevant topics of concern. According to Cronbach and Meehl

(1955), the construct is some postulated attribute of people, assumed to be reflected in test performance. The question can be asked, Does the instrument measure the concept, construct, or content that it's intended to measure? For example, if you are trying to measure stress does the instrument measure stress the same for all respondents regardless of ethnic background, age, country, etc.

3.3 Reliability

Reliability can be tested or estimated by comparing different versions of the same measurement or instrument. Reliability is more concerned with the ability of an instrument to measure consistently. There are several types of reliability tests including test-retest, interrater and internal consistency. In quantitative research, the most popular reliability test is internal consistency using Cronbach's alpha, often referred to as just alpha. Calculating alpha has become common practice in educational research when multiple-item measures of a concept or construct are employed (Tavakol & Dennick, 2011). Cronbach's alpha is also used as an indicator in the development of scales intended to measure attitude and other affective constructs (Taber, 2018).

Reliability is a major concern in social science and TVET research when measurement or test is used to measure some attribute of behaviour. Reliability measures the consistency or repeatability of measurement or test results. For example, one researcher used an inventory research instrument to measure the perception of respondents on soft skills using the Likert Scale. Several different researchers also used the same research instrument but in different contexts e.g., different locations. Since the construct and items in the research instrument are the same, the reliability test (internal consistency) should be the same or almost the same because it intends to measure the same items on soft skills. When different researchers perform the measurements or tests, on different occasions, under different conditions, with supposedly alternative instruments which measure the same things. If the reliability test is not the same, it means that the instruments are unreliable and should be re-evaluated before using them. Reliability can cause a problem as the majority of parametric statistical procedures assume that sample data are measured without error and the result of poor reliability might present a problem for descriptive statistics such as the mean because part of the average score is actually an error (Nimon, Zientek, & Henson, 2012). Thus, reliability is the degree to which the observed score of a measure reflects the true score of that measure. There is clearly some overlap between if an instrument's response process is valid and reliable. The important thing for the beginning researcher to think about is if their study is measuring what it is supposed to be measuring and measures that content or construct with the sample in the study.

4. Methodology

The research design and approach used for this study was a non-experimental research design. Data were using in this research are retrospective data or pre-existing data. The important feature of retrospective data is that the data is already available and don't require any data collection (Talari & Goyal, 2020). The non-experimental approach was chosen because it focuses on a statistical relationship between two variables but does not include the manipulation of an independent variable, random assignment of participants to conditions or orders of conditions, or both. The simulation uses statistical software known as Statistical Package for the Social Sciences (SPSS) to show the validity and reliability that can be done. Once the data were input and the variables were defined, all missing data were treated and cleaned using Exploratory Data Analysis (EDA), and the analysis of reliability is conducted. Through the SPSS Output, each output table is carefully examined, inspected, and interpreted. These simulating data models are used for many replications, applying statistical models to the resulting data sets to create estimates of the key statistics, and assessing the validity and reliability of these estimates to showcase the important part of the analytic process. The key benefit of using simulation is, the simulations can be used to strengthen and enhance the researcher's understanding especially novice researchers of understanding validity and reliability.

5. Findings and Discussion

Reliability tests can be assessed in different ways, for instance using test-retest reliability for stability, inter-item reliability for internal consistency, interrater reliability, or parallel scale for equivalence. However, the most used reliability estimator in social science and TVET research is Cronbach's Alpha. This test was introduced in 1951 by Cronbach, as a generalization of the KR-20 estimator created, in 1937 by Kuder and Richardson. Most of the research in social science and TVET areas used an internal consistency value as a reference to accept or reject any items in a construct. These hypothetical constructs are not directly observable and are called latent variables. In this context, most researchers write about the fact that they tend to reject items in the constructs or questionnaires using single items, especially due to the generally well-known issues with measurement reliability. Many researchers in social science and the TVET area agreed that any value above .70 is considered is consistent and can be accepted ($\alpha \geq .70$). Given the dominance of the internal consistency perspective, these simple results have serious implications.

Table 3 - Evaluating measurement reliability and validity coefficients (Gliner, Morgan & Leech, 2017)

Correlation coefficient	Support for reliability	Support for validity
+.90	Very good [a]	Strong but [d]
+.80	Good [b]	Strong but [d]
+.70	Adequate [b]	Strong but [d]
+.60	Minimal [b]	Strong
+.50	Not acceptable	Strong
+.30	Not acceptable	Medium
+.10	Not acceptable	Weak
-.10	Nor acceptable [c]	Weak [e]
-.30	Nor acceptable [c]	Medium [e]
-.50	Nor acceptable [c]	Strong [e]
>-.50	Nor acceptable [c]	Strong but [e]

Note:
[a] Useful for decisions about individual selection, placement, and so forth
[b] Useful for research, but probably not for a decision about individuals
[c] Check data for probable errors in coding or conceptualization
[d] If a validity coefficient is quite high (e.g., > +/-.70), you are probably measuring the same or very similar concepts, rather than two separate ones
[e] Criterion and convergent construct validity would be expected to produce positive correlations unless the concepts are hypothesized to be negatively related (e.g., anxiety and GPA)

The interpretation of reliability is the correlation of the test with itself as shown in Table 3. Squaring this correlation and subtracting it from 1.00 produces the index of measurement error. For example, if a test has a reliability of 0.70, there is a 0.51 error variance (random error) in the scores (0.70×0.70 = 0.49; 1.00 – 0.49 = 0.51), if the alpha value is high the error variance becomes low e.g., reliability of 0.90, there is a 0.19 error variance (0.90×0.90 = 0.81; 1.00 – 0.49 = 0.19). As the estimate of reliability increases, the fraction of a test score that is attributable to the error will decrease. It is of note that the reliability of a test reveals the effect of measurement error on the observed score of the group being tested rather than on an individual student. To calculate the effect of measurement error on the observed score of an individual student, the standard error of measurement must be calculated (Tavakol & Dennick, 2011).

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

Fig. 1 - Cronbach alpha formula

Where:
α = coefficient alpha
k = number of items
$\bar{r}$ = average inter-item correlation

Cronbach's alpha reliability coefficient normally ranges between 0 and 1 (Gliem & Gliem, 2003; Koonce & Kelly, 2014). However, there is no lower limit to the coefficient. When the coefficient of Cronbach's alpha approaches 1.0, a scale demonstrates greater internal consistency. Hence, the highest number for Cronbach's alpha coefficient is 1.00, meaning the greater the internal consistency of the items in the scale. Based upon Figure 1, where *k* is the number of items considered and r is the mean of the inter-item correlations the size of alpha is determined by both the number of items in the scale and the mean inter-item correlations.

Table 4 - Statistics Summary from SPSS output (n=36)

Summary Item Statistics	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	4.09	3.95	4.20	0.25	1.06	0.01	8
Item Variances	0.58	0.47	0.77	0.30	1.63	0.01	8
Inter-Item Correlations	0.50	0.27	0.76	0.50	2.85	0.02	8

Table 5 - Inter-items analysis from SPSS output (n=36)

Item Total Statistics	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
PSS1	28.65	16.90	0.58	0.65	0.88
PSS2	28.50	16.72	0.65	0.71	0.88
PSS3	28.70	15.70	0.63	0.53	0.88
PSS4	28.50	17.03	0.59	0.43	0.88
PSS5	28.60	15.89	0.70	0.68	0.87
PSS6	28.60	16.14	0.69	0.64	0.87
PSS7	28.75	14.86	0.81	0.74	0.86
PSS8	28.60	16.40	0.64	0.55	0.88

Table 6 - Reliability Coefficients from SPSS output (n=36)

N of Items	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items
8	0.88	0.88

Using retrospective data that are available, the researchers run a reliability test using the Cronbach Alpha's test to check on the internal consistency. Internal consistency should be determined during the pilot test and before the actual data collection can be employed for research to ensure validity. Furthermore, reliability is used to estimate and show the amount of measurement error in a test. Tables 4,5, & 6 show an example of the item-analysis output from SPSS for the multi-item scale of one construct named Problem-Solving Skills in the Transversal Skills study. Gliner, Morgan, and Leech (2009) provide the following guideline: $\alpha > 0.90$ = Very Good, $\alpha > 0.80$ = Good, $\alpha > 0.70$ = Adequate, $\alpha > 0.60$ = Minimal, and $\alpha < .50$ = Not Acceptable as shown in Table. The accepted value of Cronbach's alpha in social sciences is 0.70, however, values above 0.6 are also accepted depending on the field of study (van Griethuijsen, et al., 2015; Taber, 2018). The reliability coefficient for this pilot study is 0.88 suggests that it is good, and it indicates strong internal consistency among the eight items. However, if the coefficient value is too high (over 0.90) this can be a problem of multicollinearity. In other words, one predictor variable can be used to predict the other. This creates redundant information, skewing the results. Examples of correlated predictor variables also known as multicollinear predictors are when the question is asked about a person's height and weight, and the respondents think it is the same thing. Hence, reliability can be improved by writing items clearly and making test instructions easily understood by the respondents or by removing redundant items. This simulation shows that by using alpha, the researcher can determine internal consistency and measure consistency within the instrument, and question how well a set of items measures a particular behaviour or characteristic within the test.

Other than Cronbach's alpha value, the researcher is encouraged to check the inter-item corrections. It is recommended that, in an empirical approach and as a guide, if the score of the inter-item correlations exceeds 0.30, the validity of construct-related evidence or evidence-based on the internal structure is satisfied (Field, 2013). Items with very low correlations (< 0.30) are less desirable and could be a cue for potential deletion from the tentative scale (Boateng et al., 2018). Similarly, Cristobal, Flavián, and Guinalíu (2007) purported that the subscales with corrected item-total correlation lower than 0.30 are not acceptable. Furthermore, if the inter-item correlation lies between 0.10 and 0.29, then there is a weak correlation for both positive and negative values, and when the inter-item correlation lies between 0.30 and 0.49 a medium correlation, and lastly if the inter-item correlation is between 0.50 and 1.00 a strong correlation (Cohen, 1988).

Table 7 - Inter-Item Correlation Matrix from SPSS output (n=36)

	PSS1	PSS2	PSS3	PSS4	PSS5	PSS6	PSS7	PSS8
PSS1	1.000	.763	.491	.502	.268	.280	.514	.328
PSS2	.763	1.000	.638	.402	.346	.361	.591	.311
PSS3	.491	.638	1.000	.298	.489	.432	.518	.471
PSS4	.502	.402	.298	1.000	.489	.461	.547	.461
PSS5	.268	.346	.489	.489	1.000	.691	.748	.602
PSS6	.280	.361	.432	.461	.691	1.000	.701	.676
PSS7	.514	.591	.518	.547	.748	.701	1.000	.538
PSS8	.328	.311	.471	.461	.602	.676	.538	1.000

While increasing the value of alpha is partially dependent upon the number of items on the scale, it should be noted that this has diminishing returns. That is, you do not want your construct being measured to have more items than necessary. For example, if you can reduce the length of your instrument from nine questions to five, and keep similar validity and reliability measures, then it is wise to do so. It should also be noted that an alpha of 0.80 is probably a reasonable goal. Also, while a high value for Cronbach's alpha indicates good internal consistency of the items in the scale, it does not mean that the scale is unidimensional. Factor analysis is a method to determine the dimensionality of a scale and should be considered. Factor analysis also can be used to investigate the structure and validity of items in a study or research (Salleh, Sulaiman, & Gloeckner, 2015). Other than that, the researcher also should consider other threats. There are many threats to the validity and reliability of a research design and testing. Some of these threats include history, maturation, testing, instrumentation, selection, mortality, diffusion of treatment and compensatory equalization, rivalry, demoralization, and others. Most important, the researcher should be very careful in making conclusions because reliability on its own is not enough to ensure validity. Even if a test is reliable, it may not accurately reflect the real situation.

6. Conclusion

This paper provides a simulation and detailed explanation of how to use and conduct validity and reliability in social science and TVET research, especially for novice researchers. The method used in this research is computer simulation in SPSS. Using retrospective data or pre-existing data, the authors were able to simulate the result which help explain reliability and validity. Based on the result from this simulation it indicates the importance of conducting validity and reliability test during the research process. It not only provided evidence but also confidence because it reflected consistency and accuracy. While reliability is important for study, it is not sufficient unless it is combined with validity. Hence, for a test or research instrument to be reliable, it also needs to be valid. A research agenda for validity and reliability should be a priority for social sciences assessments, especially in education and TVET. In the end, this paper intended to provide insight, knowledge, and critical argument on the important concept of validity and reliability. This also allows the novice researcher to understand some basic concepts of measuring validity and reliability but also hopefully stimulates interest in additional detailed measures.

Acknowledgement

This research was made possible by funding from research grant number K373 provided by the Ministry of Higher Education, Malaysia. The authors would also like to thank the Faculty of Universiti Tun Hussein Onn Malaysia for its support.

References

American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME] (2014). *The standards for educational and psychological testing.* AERA-APA-NCME.

Baimyrzaeva, M., (2018). *Beginners' guide for applied research process: What is it, and why and how to do it.* Kyrgyzstan: University of Central Asia, pp.10-26.

Benitez, I., & Padilla, J. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*(149), 1-18.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin, 52,* 281-302.

Cristobal, E., Flavián, C., & Guinalíu, M. (2007). Perceived e-service quality (PeSQ): Measurement validation and effects on consumer satisfaction and website loyalty. *Managing Service Quality: An International J.,17*(3), 317-340.

Drost, E. A. (2011). Validity and reliability in social science research. *Edu. Research and Perspective, 38*(1), 105-124.

Faulkner-Bond, M., & Sireci, S. (2013). Validity evidence based on test content. *Psicothema, 26*(1), 100-107.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Gliner, J. A., Morgan, G. A., & Leech, N. L. (2017). *Research methods in applied settings: An integrated approach to design and analysis* (3rd ed.). Routledge.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert-type scales. *2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education, Columbus*, 82-88.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181-191.

Haradhan, M. (2017). Two criteria for good measurements in research: Validity and Reliability. *Annals of Spiru Haret University, 17*(3), 58-82.

Heale, R. & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evid Based Nurs, 18*(3), 66-67.

Koo, T. K. & L, M. Y, (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Koonce, G. L., & Kelly, M. D. (2014). Analysis of the reliability and validity of a mentor's assessment for principal internships. *NCPEA Education Leadership Review, 15*(2), 33-48.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127-135.

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3,* 102.

van Griethuijsen, R. A. L. F., van Eijck, M. W., Haste, H., Brok, N. C. S., Mansour, N., Gencer, A. S, & BouJaude, S. (2015). Global patterns in students' views of science and interest in science. *Research in Science Edu., 45,* 581-603.

Salleh, K. M., Sulaiman, N. L., & Gloeckner, G. (2015). The development of the competency model perceived by Malaysian human resource practitioners' perspectives. *Asian Social Science, 11*(10), 175-185.

Salleh, K. M. & Suliman, N. L. (2015). Technical skills evaluation based on competency model for human resource development in Technical and Vocational Education. *Asian Social Science, 11*(16), 74-79.

Salvia, J. & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Houghton Mifflin Company.

Sireci, S. & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26,* 100-107.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48,* 1273-1296.

Taherdoost, H. (2016). Validity and reliability of research instrument: How to test the validation of questionnaire/survey in research. *International Journal of Academic Research in Management, 5*(3), 28-36.

Talari, K. & Goyal, M. (2020). Retrospective studies: Utility and caveats. *Journal of the Royal College of Physicians of Edinburgh, 50*(4), 398-402.

Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach's alpha. *International J. of Medical Education, 2*, 53-55.

Wells, C.& Rios, J. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.