Mediterranean Journal of Social & Behavioral Research
2022, 6(2), 53-59
ISSN 2547-8559 (Print)
ISSN 2547-8567 (Online)
https://www.mjosbr.com/

Review Article

OPEN ACCESS

A Systematic Review of Reporting of Psychometric Properties in Educational Research

Simon Ntumi [1]*, Kwesi Twum Antwi-Agyakwa [2]

[1] Department of Educational Foundations, Faculty of Educational Studies, University of Education, Winneba, West Africa, GHANA
[2] Department of Fisheries and Aquaculture, University of Cape Coast, West Africa, GHANA
***Corresponding Author:** sntumi@uew.edu.gh

Citation: Ntumi, S., & Twum Antwi-Agyakwa, K. (2022). A Systematic Review of Reporting of Psychometric Properties in Educational Research. *Mediterranean Journal of Social & Behavioral Research*, 6(2), 53-59. https://doi.org/10.30935/mjosbr/11912

ABSTRACT

Background: The validity and reliability of research outputs are important elements of the research trail. They drive accuracy, transparency, and minimize researcher biases, contributing to rigor and dependability. This paper reviews the frequency of published articles reporting the psychometric properties of the scales/subscales employed in educational research.

Methods: We conducted a systematic review of psychometric properties in educational research papers published between 2010 and 2020 from 15 education-related journals. In our search, we included quantitative studies with primary data. The methodological quality assessment was performed using trained reviewers. The search was conducted using PRISMA 2020 to identify, screen eligible papers for inclusion. The extracted was analyzed using SPSS v25 while reported and interpreted in descriptive statistics.

Findings: We extracted 763 papers published between 2010 and 2020 from 15 education-related journals. More than half of the articles reviewed did not report either validity (n=456 out of 763, 59.8%) or reliability (n=400, out of 763, 52.4%) statistic. For those reporting either validity or reliability, the alpha coefficient was the most widely used statistic to establish reliability (n=185, 50.9%) and correlation coefficient was frequently reported (n=219, 71.3%) for validity.

Conclusions: The paper concluded that to produce dependable conclusions and recommendations in educational research, it is imperative for researchers to pursue psychometric properties to ground their findings and take-home learning.

Keywords: validity, reliability, educational research, psychometric properties

Received: 31 Jan. 2022 ◆ Accepted: 18 Mar. 2022

INTRODUCTION

Educational measurement using validity and reliability plays a crucial role in the social sciences (Barry et al., 2014; Mohajan, 2017). Specifically, educational researchers and practitioners often develop, adapt, or adopt surveys/scales to quantify and measure pertinent participant characteristics (e.g., cognitive, behavioral, emotional, classroom factors, assessment issues, and psychological factors). It is therefore vital for one to note that the integrity of these measurements is critical to the derivation of sound research conclusions (Barry et al., 2014; Bull et al., 2019). Nevertheless, in order to draw accurate conclusions based on survey data or scale, there is the need for a certain level of expertise. Two issues, in particular, are intrinsically tied to interpreting measurement results from surveys: validity and reliability (Corbett et al., 2015; Hogan & Agnello, 2004).

In educational research and allied field (e.g like psychology, nursing, and counselling), it is asserted that validity and reliability are the two most important and fundamental features in the evaluation of any measurement instrument or tool for quality research (Kimberlin & Winterstein, 2008). In essence, the evidence of validity and reliability are rudiments to assure the integrity and quality of a measuring instrument (Flake et al., 2017; Kimberlin & Winterstein, 2008). Forza (2002) adds that without assessing reliability and validity of research instrument, it will be difficult to describe the extent of measurement errors and ascertain any theoretical relationships among the concepts being studied.

Validity is generally described as the extent to which an instrument measures what it asserts to measure (Blumberg et al., 2005; Plake & Wise, 2014). Put differently, validity is the degree to which the results are 'truthful'. Validity allows us to establish whether the results obtained meet all of the requirements of the scientific research method. Indeed, some scholars view validity as a "compulsory" requirement of the scientific endeavour (Oliver, 2010). These descriptions more mimics the validation process in quantitative studies. In qualitative research, validity is seen as trustworthiness, utility, and dependability (Liang et al., 2014; Zohrabi, 2013). In this context, validity connotes scrupulous compliance to a particular research paradigm (e.g., grounded theory and phenomenology) during the process of generating research findings.

On the other hand, reliability can be explained as a measurement that supplies consistent results in several occasions (Blumberg et al., 2005; Twycross & Shields, 2004). Reliability measures consistency, precision, repeatability, and trustworthiness of a research (Campos et al., 2017; Chakrabartty, 2013; Squires et al., 2011; Yarnold, 2014). It indicates the extent to which it is without bias (error free); and hence, insures consistent measurement across time and the items in an instrument (the observed scores). Some qualitative researchers use the term 'dependability' instead of reliability.

Considering the multidimensionality nature of research and its implications, it is important that issues of how psychometric properties in educational research are reported in studies should be given the needed attention. In this regard, a study to reveal the significant of psychometric properties in studies appears very relevant and needed to measure the accuracy of research findings. To this end, it becomes necessary to conduct a systematic review for evaluating the psychometric properties of scales/instruments that measure the accuracy, transparency and dependability of research findings published on research journals. Specifically, the study rides on finding out the frequency of validity and reliability reporting practices by author in the selected publication houses, also the study sought to assessed the most frequently reported types of validity and associated statistics in studies and finally, the study examined the frequency of reported types of reliability and associated statistics in educational related studies.

METHODS

Study Selection (Inclusion & Exclusion Criteria)

The Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA) 2020 statement guided the methodology and reporting of this systematic review. 15 journals in educational research from five publishing houses were sampled: SAGE Publications, Springer, Elsevier, Multidisciplinary Digital Publishing Institute (MDPI) Journals, and Francis & Taylor. The data were extracted from published articles only from the five selected publication houses. We searched the method sections of the accessed papers that guided the inclusion criteria. The inclusion criteria were based on quantitative study, collected primary data, and published between 2010 and 2020. Mixed method studies were included, but only the quantitative portion was examined for this investigation. However, we excluded letters to editor, commentaries, and conceptual and/or theoretical studies. Each of the author independently reviewed the papers that were included in the final sample.

The selected studies involved human subjects, and collected primary data on experience, perpetration, or response to educational issues across the globe. Furthermore, we included observational studies (e.g., cross-sectional studies, cohort studies, and case-control studies). We were guided by how validity and reliability are used as a psychometric property in reporting. We assessed relevance based on title and abstract. Secondary reviewers randomly conducted relevancy check for 10% of studies we (primary reviewers) considered "cannot determine". The discrepancies on relevancy of the articles between primary and secondary reviewer were noted and discussed by the entire team and consensus agreement was reached.

Data Extraction, Quality Assessment & Data Analysis Procedure

To extract the data, the eligible articles went through a standardized data extraction and quality assessment process. The data extraction form was refined during the extraction of the first few articles to ensure that the forms were comprehensive. We extracted descriptive characteristics of the sample from each quantitative. Extracted data from eligible studies were compiled using the guidelines of PRISMA. To ascertain this, each reviewer assessed to find out if articles provided validity and reliability statistics in the analysis or methodology or in the literature. Again, the reviewers checked if the statistics or the psychometric properties were from a previous administration of the instrument or the current sample. Again, we assessed the authenticity of the validity and reliability statistic and how they were reported or estimated by authors in various studies. Consequently, the third reviewer compiled results by importing data into the Statistical Package for Social Sciences (SPSS) v25 and conducted data validation for data entry and coding errors. This was done by recording all the string variables and verifying accuracy of all the entered data. How the data were extracted from the publications house is represented in the PRISMA flow chat in **Figure 1**. **Figure 1** presents the PRISMA flow chat of how the data was extracted from the journals database.

These quality assessment methods by collective reviewers helped in ensuring the reduction of bias by gaining some level of accuracy and transparency. The results reported in this study are based on a final sample (n=763) published articles from 15 journals from five publication houses. This sample is a comprehensive representation of the articles that are related to educational research, straddling a total of 25 volumes across 10 distinct years (2010-2020). The obtained data were cleaned and processed using the SPSS v25 and analyzed using descriptive statistics (frequencies and percentages).

FINDINGS

The analysis of the paper of the was based on 763 published articles gathered from five major publication houses (this included: SAGE Publications, Springer, Elsevier, MDPI, and Francis & Taylor). **Table 1** presents the results of number of articles (n=763) and validity and reliability reporting by the selected publication companies. The results present the analysis under three themes. The reviewers looked at the number of articles and validity and reliability reporting by the selected journals (**Table 1**). As presented in **Table 1**, generally, the results suggest that most of the articles published in the five (n=5) selected publications houses were not reporting either validity (n=456, 59.8%) or reliability (n=400, 52.4%). For example, for studies published in SAGE, the results show that, most of the articles were not reporting validity (n=95, 60.1% out of a sample 158). Those who reported validity were few (n=63, 39.9% out of a sample of 158). On reliability, it was evident that, majority of the articles were not placing much emphasis reliability in their studies (n=83, 52.5%, out of a sample of 158).

Similarly, from the Springer, it was found that more than half of the articles were not validity in their studies (n=110, 58.2%, out of a sample of 189). On reliability, contrary evidences were recounted as it was found that mores half (n=105, 55.6%, out of a sample of 189) of the papers published in the Springer reported reliability in their study. Ultimately, this percentage recorded in Springer articles did not have any significant impact when all the papers are reviewed holistically.

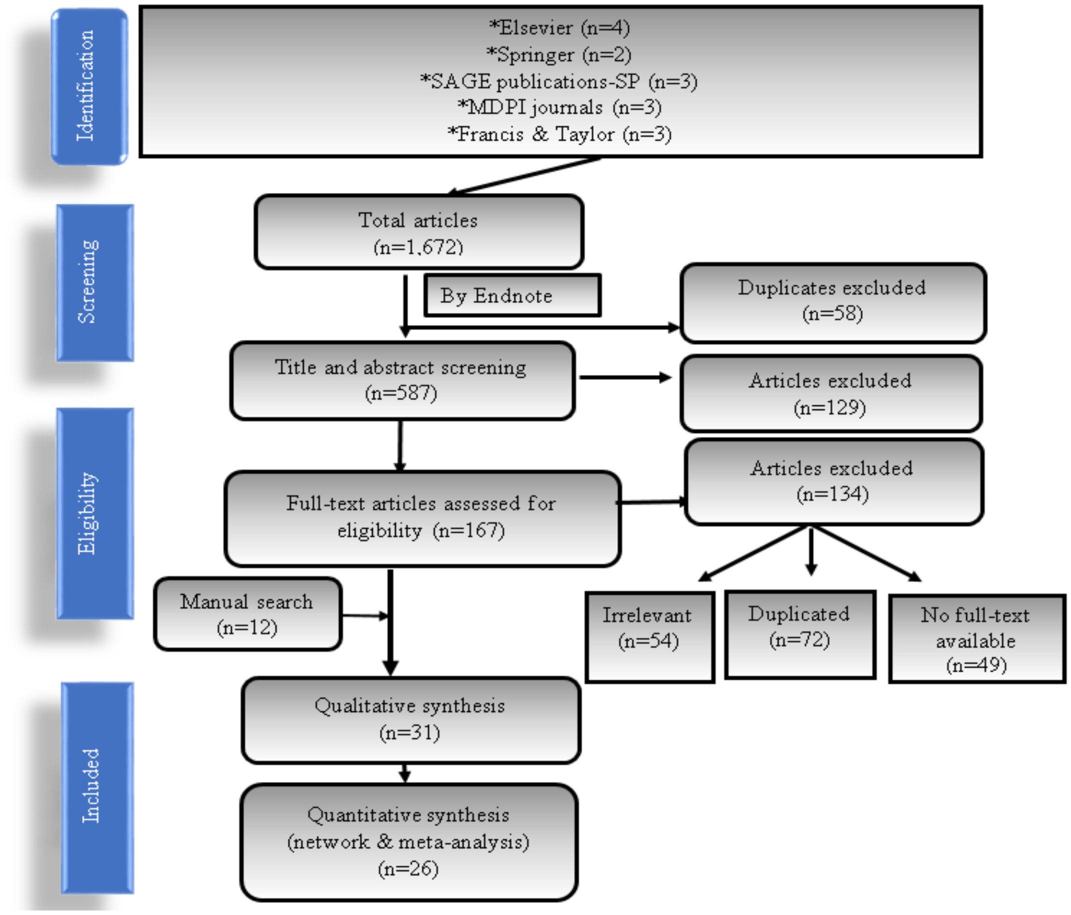
Identification
*Elsevier (n=4)
*Springer (n=2)
*SAGE publications-SP (n=3)
*MDPI journals (n=3)
*Francis & Taylor (n=3)
Screening
Total articles (n=1,672)
By Endnote
Duplicates excluded (n=58)
Title and abstract screening (n=587)
Articles excluded (n=129)
Eligibility
Articles excluded (n=134)
Full-text articles assessed for eligibility (n=167)
Manual search (n=12)
Irrelevant (n=54)
Duplicated (n=72)
No full-text available (n=49)
Included
Qualitative synthesis (n=31)
Quantitative synthesis (network & meta-analysis) (n=26)

Figure 1. PRISMA 2020 detailed flow diagram guideline for systematic review steps

Table 1. Frequency of validity & reliability reporting practices by authors in the selected publication houses

		Validity (v)		**Reliability (r)**	
Publications houses	**Total sample (n, %)**	**Reported (n, %)**	**Not reported (n, %)**	**Reported (n, %)**	**Not reported (n, %)**
SAGE publications	(n=158, 20.7%)	(n=63, 39.9%, <50	(n=95, 60.1%, >50)	(n=75, 47.5%, <50)	(n=83, 52.5%, >50)
Springer	(n=189, 24.8%)	(n=79, 41.7%, <50)	(n=110, 58.2%, >50)	(n=105, 55.6%, >50)	(n=84, 44.4%, <50)
Elsevier	(n=132, 17.3%)	(n=46, 34.8%, <50)	(n=86, 65.1%, >50)	(n=62, 46.9%, <50)	(n=70, 53.1%, >50)
MDPI journals	(n=149, 19.5%)	(n=58, 38.9%, <50)	(n=91, 61.1%, >50)	(n=57, 38.1%, <50)	(n=92, 61.7%, >50)
Francis & Taylor	(n=135, 17.7%)	(n=61, 45.2%, <50	(n=74, 54.8%, >50)	(n=64, 47.4%, <50)	(n=71, 52.6%, >50)
Total	763	**(n=307, 40.2%)**	**(n=456, 59.8%)**	**(n=363, 47.6%)**	**(n=400, 52.4%)**

Note. Analysis of extracted data from publication house database, Key-n=sample, %-percentage

With respect to studies published in Elsevier, it was found that majority of the reviewed articles did not report on validity (n=86, 65.1%, out of a sample of 132). Similar findings were accrued in respect to the review on reliability as it was found that more than half of the papers (n=70, 53.1%, out of a sample of 132) from Elsevier did not report reliability in their study. For studies published in MDPI, it was found that majority of the papers (n=74, 54.8%, out of a sample of 149) were silent on validity. Comparable results were recounted on the issues of reliability as most of the papers from MDPI journals failed to report on reliability (n=92, 61.7%, out of a sample of 149). Finally, it was found that most of the authors publishing in Francis & Taylor failed to report on validity in their studies (n=74, 54.8%, out of a sample of 135). On reliability, similar evidence was established on reliability as more than half of the papers were stillness on reliability (n=71, 52.6%, out of a sample of 135).

Table 2 depicts the type of validity components reported in the selected articles. The components were categorized under content construct, face, predictive, criterion, and multiple types of validity.

From **Table 2**, the results suggest that most of the most of the articles were falling on pilot testing to report the validity results (n=124, 40.4% out of 307 articles). Among all the sub-components under content validity. Cognitive interviews were least reported (n=37, 12.1% out of 307 articles). Dwelling on the construct validity, it was found that most of the papers focused on correlation coefficient to report their validity results (n=159, 51.0% out of 307 articles). Chi-square test as a means to report validity was the least (n=53, 17.3% out of 307 articles).

For those works that reported on face validity, it was found that more than half of the articles were using expert panel assessment to validate their instruments (n=209, 68.1% out of 307 articles). Those which used factor analysis were the slightest (n=48, 15.6% out of 307 articles). We again reviewed on predictive validity and the results showed that most of the articles were concentrating on correlation coefficient to report their predictive validity (n=219, 71.3% out of 307 articles). Logistic regression was less used in the articles to report on predictive validity (n=90, 29.3% out of 307 articles).

Table 2. Most frequently reported validity types & associated statistics

Validity components	n	%	Rank order
Content validity			
Expert panel assessment	58	18.9	3rd
Pilot testing	124	40.4	1st
Literature review	88	28.7	2nd
Cognitive interviews	37	12.1	4th
Construct validity			
Factor analysis	95	31.0	2nd
Correlation coefficient	159	51.0	1st
Chi square	53	17.3	3rd
Face validity			
Correlation coefficient	50	16.3	2nd
Expert panel assessment	209	68.1	1st
Factor analysis	48	15.6	3rd
Predictive validity			
Correlation coefficient	219	71.3	1st
Logistic regression	90	29.3	2nd
Criterion validity			
Correlation coefficient	199	64.8	1st
Factor analysis	108	35.2	2nd
Multiple types of validity reported			
Content & face validity	102	33.2	1st
Expert panel assessment & literature	98	31.9	2nd
Content & construct validity	57	18.6	3rd
Factor analysis & logistics	50	16.3	4th

Note. n=**307** (307 is based on the confirmed reported validity results in **Table 1**)

The next component that was assessed and reviewed is criterion validity. In this component, it was found that correlation coefficient was highly used to report on validity results in the articles (n=199, 64.8% out of 307 articles). Generally, factor analysis is least reported in almost all the articles (n=108, 35.2% out of 307 articles). In looking at the combine effects (multiple types of validity reported), it was appreciated that most of the authors used or combine content and face validity in their report (n=102, 33.2% out of 307 articles). Factor analysis and logistics was least reported in all reviewed articles (n=50, 16.3% out of 307 articles).

Finally, the most frequently reported types of reliability and associated statistic is reported in **Table 3**. In **Table 3**, the reviewed was guided by five associated statistics for obtaining reliability. These are internal consistency, interobserver/interrater, test-retest, parallel form of reliability and multiple (combination) types of reliability reported. Reviewing on internal consistency, it was found that most of the articles were reporting on Alpha coefficient to establish their reliability (n=185, 50.9%, out of a sample of 363). Kappa coefficient was least reported (n=24, 6.61%, out of a sample of 363).

Mirroring to how interobserver/interrater was reported in studies, it was found that Alpha coefficient was highly reported (n=298, 81.5%, out of a sample of 363). Kappa coefficient (n=65, 17.9 %, out of a sample of 363) was least reported. In relation to test-retest, it was revealed that alpha coefficient was dominating in all the articles 763 across the five (5) selected publication houses (n=196, 53.9%, out of a sample of 363). In terms of test-retest, kappa coefficient was least reported in the studies (n=53, 14.6%, out of a sample of 363).

In assessing the parallel form of reliability, it was found that most of the articles were reporting Alpha coefficient when the authors want to establish parallel form of reliability (n=223, 61.4%, out of a sample of 363). In furtherance to the above, it was found that correlation coefficient was least reported in the articles (n=140, 38.6%, out of a sample of 363). In the combination of the associated statistics, it was found that most of the authors syndicate correlation coefficient and Alpha coefficient to estimated their reliability (n=313, 86.2%, out of a sample of 363). Very few of them were merging Alpha coefficient and kappa coefficient to estimate or determine their reliability statistic (n=29, 7.99%, out of a sample of 363).

Table 3. Most frequently reported reliability types & associated statistics

Reported reliability type	n	%	Rank order
Internal consistency			
Alpha coefficient	185	50.9	1st
Correlation coefficient	154	42.2	2nd
Kappa coefficient	24	6.61	3rd
Interobserver/interrater			
Kappa coefficient	65	17.9	2nd
Alpha coefficient	298	81.5	1st
Test/re-test			
Alpha coefficient	196	53.9	1st
Correlation coefficient	114	31.4	2nd
Kappa coefficient	53	14.6	3rd
Parallel form of reliability			
Alpha coefficient	223	61.4	1st
Correlation coefficient	140	38.6	2nd
Multiple (combination) types of reliability reported			
Correlation & alpha coefficient	313	86.2	1st
Kappa & correlation coefficient	21	5.79	3rd
Alpha coefficient & Kappa coefficient	29	7.99	2nd

Note. n=**363** (363 is based on the confirmed reported validity results in **Table 1**)

DISCUSSION

The study reviewed psychometric properties reporting practices among published studies in educational related fields. Riding on the work of Bannigan and Watson (2009), it was opined that at one level, the concepts of reliability and validity are relatively easy to understand nevertheless, when it comes to translating this into the reality of psychometrics in research, where concepts are not tangible and standards are scarce then it becomes less easy to understand and, in fact, quite complex. Based on the postulation from Bannigan and Watson (2009), the study tried to determine the frequency with which published articles appearing in high impact journals report the psychometric properties of the scales/subscales employed and also, to outline the methods or associated statistics to determine the reliability and validity estimate or statistic.

The study revealed that the most published article in education discipline appear not to give the needed attention to validity and reliability. Explicitly, it was evident that most scholarly articles are failing to report either validity (n=456 out of 763, 59.8%) or reliability (n=400, out of 763, 52.4%) statistics. Again, for those reporting either validity or reliability, the Alpha coefficient was the most widely used statistic to establish reliability (n=185, 50.9%) and correlation coefficient was frequently reported (n=219, 71.3%) for validity.

The evidences obtained from the study lend empirical support to related works on how psychometric properties are reported in studies. To be specific, in the work of Barry et al. (2011), they asserted that researchers have frequently noted the need of assessing and reporting measures of reliability and validity with each administration of a survey/scale. However, most researchers recurrently fail to acknowledge the psychometric properties of validity and reliability in their studies. Again, the results from this study share similar findings with the study of Adams et al. (2014) and Singh (2014), which reported in their study that out of 967 published articles, spanning seven prominent health education and behavior journals between 2007 and 2010, an exceedingly high percentage failed to report either validity (ranging from 40% to 93%) or reliability (ranging from 35% to 80%) statistics in the selected studies.

In furtherance to the above, the results from this study concur with the work of Squires et al. (2011) who found in their study that reliability was least reported (33%) in many studies. They further asserted that internal consistency (Cronbach's alpha) reliability was reported in 31 studies; values exceeded 0.70 in 29 studies. Their study further indicated that test-retest reliability was reported in three studies with Pearson's r coefficients >0.80. No validity information was reported for 12 of the 60 measures. According to their study, the remaining 48 measures were classified into a three-level validity hierarchy according to the number of validity sources reported in 50% or more of the studies using the measure. Level one measures (n=6) reported evidence from any three (out of four possible) standards validity sources (which, in the case of single item measures, was all applicable validity sources). Level two measures (n=16) had evidence from any two validity sources, and level three measures (n=26) from only one validity source.

In some previous review like the work of Bolarinwa (2015), Estabrooks et al. (2003), and Tavakol and Dennick (2011), similar evidences were accrued. These studies were conducted within the same umbrella of finding out the authors report psychometric properties in their studies. Generally, these studies found that most studies lack significant psychometric assessment of used instruments. These studies further stated that over half of the studies in their review did not mention validity or reliability in their report.

Relatedly, we see that oversight (either internal or not internal) is very troubling and disturbing in research works. In essence, most researchers failing to report these important psychometric properties, suggest that most of the researchers might be making or reporting erroneous conclusions and recommendations. In other words, Barry et al. (2014) put it this way that by not ensuring the instruments employed in a given study were able to produce accurate and consistent scores, researchers cannot be certain they actually measured the behaviours and/or constructs reported. To this end, it is highly possible that researchers may be unknowingly measuring something entirely different construct than originally intended construct to be measured (Adams et al., 2014; Moana-Filho et al., 2017).

The results further corroborate with the work of Mahmood (2017). The purpose of Mahmood's (2017) was to present the results of a review of the evidence on psychometric properties of information literacy tests. The study found that the most commonly used psychometric analysis included content validity, discrimination validity and internal consistent reliability. Similar to the present study, it was found that Alpha coefficient was the most widely used statistic to establish reliability (n=185, 50.9%) and correlation coefficient was frequently reported (n=219, 71.3%) for validity.

Relatedly, our study shares common findings with the study of Bull et al. (2019). The main purpose of their study was to identify Patient-Reported Experience Measures (PREMs), assess the reporting nature of validity and reliability, and assess any bias in the study design of PREM validity and reliability testing. The study found that priority was given to some psychometric properties than other. For example, internal consistency (n=58, 65.2%), structural validity (n=49, 55.1%), and content validity (n=34, 38.2%) were the most frequently reported validity and reliability tests in the sampled studies.

Inferring from the Standards for Educational and Psychological Testing (SEPT, 2014) for reporting validity and reliability in research, it is asserted that researchers should set forth clearly how instrument scores are intended to be interpreted and consequently used in their studies. The population (s) for which an instrument is intended should be delimited clearly, and the construct or constructs that the instrument is intended to assess should be described clearly to readers in studies. In reporting the psychometric in studies, the standard postulates that statements about validity should refer to particular interpretations and consequent uses. The work of Johnson et al. (2017) and Robson (2011) also gave credence to these characteristics of reporting validity and reliability in studies.

Although, in the study of Barry et al. (2014), it was argued that reporting validity measures is less essential when employing well-known scales that have been thoroughly tested with similar populations previously, they further counted the argument that it is always the best practice to report validity with newly created scales. The study of Adams et al. (2014), Barry et al. (2014), Hall et al. (1988), Hogan and Agnello (2004), Linn (2011), and Mohamad et al. (2015) reported similar findings in their reviews. Most of these studies asserted that consistently, many authors fail to give priority to psychometric properties in their studies. Common among the studies, it was concluded that even studies that report psychometric properties are only limiting themselves to Alpha coefficient/internal consistency reliability and correlation coefficient for validity.

Another striking and revealing findings of this study was when we noticed the disproportionate reporting practices between validity and reliability statistics in the reviewed articles. In the studies where psychometric properties were reported, there was disparities among the twice concepts. Interestingly, most of the articles appeared to have exhibited a tendency to report reliability over validity. This finding strongly supports a by study of Adams et al. (2014) who averred that a vast majority of the 967 articles exhibited a propensity to report reliability over validity.

Limitations

In this study, although, rigorous and comprehensive methods were used for the review, there are some study limitations that need to be reported. In the first place, while we reviewed articles, we did not search all grey literature sources and this might have accrued to some elements of bias. Also, our decision to exclude articles that are not related to education may be responsible for the limited number of articles sampling reporting validity and reliability. Again, studies published in other languages other than English were excluded from the review. Consequently, there may have been some relevant findings regarding the reporting practices of psychometric properties that are not captured in this review.

Finally, authors of included studies were not contacted therefore some information regarding psychometric properties of their studies may have been overlooked. To this end, a future review examining the psychometric properties of self-report measures that covers many discipline and languages would therefore be a fruitful avenue of inquiry.

CONCLUSIONS

Guided by literature, we must emphasize that a systematic review involves a critical and reproducible summary of the results of the available publications on a particular topic or clinical question. In all, the accumulated findings from the review suggest that most authors even though publishing in high profile journals appears to be relegating validity and reliability to the background. This gives the impression that current practices of not reporting validity and reliability does not mirror recommended testing practices in education research. From the review, we could infer that practice of reporting psychometric properties (validity and reliability) is currently underrepresented in the literature. In this regard, our findings point to the need for reporting of psychometric properties to be explicitly outlined as a requirement for publications. This will ensure that the practice becomes conventional for educational researchers publishing works that are related to educational measurement. To this end, it is essential to note that for researchers in education to maximize the utility and applicability of their findings for theory and practice, they must appraise and estimate the psychometric properties of their instrument employed. The neglect to this advice on psychometric properties of validity and reliability suggest that it is possible that the overall efforts of many researchers will be in vain. This is to mean that limited funds may be unnecessarily wasted on studies that are not valid and for that matter not reliable, and in essence, policies and decisions may be informed by inaccurate data and recommendations.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approve final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Acknowledgements: The authors would like to thank to anonymous reviewer for the intellectual encouragement and constructive comments throughout the manuscript. The authors would like to thank to the research assistants for their dedication during data extraction and screening process.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analysed during this study are available from the authors on request.

REFERENCES

Adams, E. J., Goad, M., Sahlqvist, S., Bull, F. C., Cooper, A. R., Ogilvie, D., & Connect, C. (2014). Reliability and validity of the transport and physical activity questionnaire (TPAQ) for assessing physical activity behaviour. *PloS One, 9*(9), 107-139. https://doi.org/10.1371/journal.pone.0107039

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing, 18*(23), 3237-3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behaviour: A review of seven journals. *Health Education & Behaviour, 41*(1), 12-18. https://doi.org/10.1177/1090198113483139

Blumberg, E. J., Hovell, M. F., Kelley, N. J., Vera, A. Y., Sipan, C. L., & Berg, J. P. (2005). Self-report INH adherence measures were reliable and valid in Latino adolescents with latent tuberculosis infection. *Journal of Clinical Epidemiology, 58*(6), 645-648. https://doi.org/10.1016/j.jclinepi.2004.11.022

Bolarinwa, O. A. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal, 22*(4), 195-101. https://doi.org/10.4103/1117-1936.173959

Bull, C., Byrnes, J., Hettiarachchi, R., & Downes, M. (2019). A systematic review of the validity and reliability of patient-reported experience measures. *Health Services Research, 54*(5), 1023-1035. https://doi.org/10.1111/1475-6773.13187

Campos, C.M.C., da Silva Oliveira, D., Feitoza, A. H. P., & Cattuzzo, M. T. (2017). Reliability and content validity of the organized physical activity questionnaire for adolescents. *Educational Research, 8*(2), 21-26. https://doi.org/10.14303/er.2017.024

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education, 3*(1), 1-8. https://doi.org/10.9790/7388-0310108

Corbett, N., Sibbald, R., Stockton, P., & Wilson, A. (2015). Gross error detection: Maximising the use of data with Uba on global producer III (Part 2). In *Proceedings of the 33rd International North Sea Flow Measurement Workshop*. Tonsberg, Norway.

Estabrooks, C., Wallin, L., & Milner, M. (2003). Measuring knowledge utilization in health care. *Nursing Leadership, 3*(3),45-67.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 1-9. https://doi.org/10.1177/1948550617693063

Forza, C. (2002). Survey research in operations management: A process-based perspective. *International Journal of Operations and Production Management, 22*(2), 152-194. https://doi.org/10.1108/01443570210414310

Hall, B. W., Ward, A. W., & Comer, C. B. (1988). Published educational research: An empirical study of its quality. *Journal of Educational Research, 8*(1), 182-189. https://doi.org/10.1080/00220671.1988.10885820

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Education and Psychological Measurement, 6*(4), 802-812. https://doi.org/10.1177/0013164404264120

Johnson, R. E., Kording, K. P., Hargrove, L. J., & Sensinger, J. W. (2017). Adaptation to random and systematic errors: comparison of amputee and non-amputee control interfaces with varying levels of process noise. *PLoS One, 12*(3), 17-27. https://doi.org/10.1371/journal.pone.0170473

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacists, 65*(1), 2276-2284. https://doi.org/10.2146/ajhp070364

Liang, Y., Lau, P. W., Huang, W. Y., Maddison, R., & Baranowski, T. (2014). Validity and reliability of questionnaires measuring physical activity self-efficacy, enjoyment, social support among Hong Kong Chinese children. *Preventive Medicine Reports, 1*(2), 48-52. https://doi.org/10.1016/j.pmedr.2014.09.005

Linn, R. L. (2011). The standards for educational and psychological testing: Guidance in test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 41-52). Routledge. https://doi.org/10.4324/9780203874776-7

Mahmood, K. (2017). A systematic review of evidence on psychometric properties of information literacy tests. *Library Review, 7*(2), 20-29. https://doi.org/10.1108/LR-02-2017-0015

Moana-Filho, E. J., Alonso, A. A., Kapos, F. P., Leon-Salazar, V., Gurand, S. H., Hodges, J. S., & Nixdorf, D. R. (2017). Multifactorial assessment of measurement errors affecting intraoral quantitative sensory testing reliability. *Scandinavian Journal of Pain, 16*(6), 93-98. https://doi.org/10.1016/j.sjpain.2017.03.007

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University, 17*(4), 59-82. https://doi.org/10.26458/1746

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia-Social and Behavioural Sciences, 204*(2), 164-171. https://doi.org/10.1016/j.sbspro.2015.08.129

Oliver, V. (2010). *301 smart answers to tough business etiquette questions.* Skyhorse Publishing.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement: Issues and Practice, 33*(4), 4-12. https://doi.org/10.1111/emip.12045

Robson, C. (2011). *Real world research: A resource for users of social research methods in applied settings.* John Wiley & Sons.

Singh, A. S. (2014). Conducting case study research in non-profit organisations. *Qualitative Market Research: An International Journal, 1*(7), 77-84. https://doi.org/10.1108/QMR-04-2013-0024

Squires, J. E., Estabrooks, C. A., O'Rourke, H. M., Gustavsson, P., Newburn-Cook, C. V., & Wallin, L. (2011). A systematic review of the psychometric properties of self-report research utilization measures used in healthcare. *Implementation Science, 6*(1), 1-18. https://doi.org/10.1186/1748-5908-6-83

Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. *American Educational Research Association.* https://www.apa.org/science/programs/testing/standards

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*(7), 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Twycross, A., & Shields, L. (2004). Validity and reliability what's it all about? Part 1 validity in quantitative studies: this is one of a series of short papers on aspects of research by Alison Twycross and Linda Shields. *Paediatric Nursing, 16*(9), 28-29. https://doi.org/10.7748/paed2004.11.16.9.28.c954

Yarnold, P. R. (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: emergency severity index (version 3) and Canadian triage acuity scale. *Optimal Data Analysis, 3*(4), 50-54.

Zohrabi, M. (2013). Mixed method research: Instruments, validity, reliability and reporting findings. *Theory & Practice in Language Studies, 3*(2), 12-18. https://doi.org/10.4304/tpls.3.2.254-262

❖❖❖