Is this a good questionnaire? Dimensionality and category functioning of questionnaires used in nursing research

Abstract

Questionnaires are perhaps the most widely used measuring tools in nursing research as many studies conducted by nurses focus on understanding the underlying complex factors that are amenable to questionnaires. However, most questionnaires used for research purposes in nursing continue to display inadequate evidence of validity under the traditional methods while ignoring the modern Rasch techniques with better proofs of objective measurement. For questionnaire data to be suitable for statistical analysis, transparent demonstration of mathematical assumptions ratified in the questionnaire are compulsory. The failure to engage contemporary measurement models in designing good questionnaires raises concerns about researchers' awareness of the application and usefulness of the evidence generated by the modern approach. The overall objective here is to draw researchers' attention to the recurrent limitations of classical approach to questionnaire design and to suggest advanced psychometric analysis exemplified in Rasch methodology as a more appropriate alternative. Therefore, this paper illustrates with examples the problems inherent in the traditional or classical test theory and advanced dimensionality and category functioning of a questionnaire as requisite psychometric properties of a questionnaire. In straightforward language, a unidimensional questionnaire evaluates a single research variable or construct per set of indicators or items on a questionnaire. Also, category functioning is an assumption test that examines the possibility that participants may fail to conserve the hierarchical order of the questionnaire categories in the way they have responded to the questions. Finally, a number of diagnostic parameters Rasch proponents recommended for testing the two essential validity criteria are outlined.

Background:

Accurate measurement of variables in quantitative research methods is central to conducting meaningful statistical analysis, drawing useful inferences, postulating empirical theories and advancing evidence-based practice in nursing and allied disciplines (Curtis & Drennan, 2013; Curtis, et al., 2016; Colgrave, et al., 2020). In clinical practice, advanced measurement science, applied in medical engineering, is the bedrock of many sophisticated pieces of medical

equipment such as blood gas analysers, electrocardiograms, pulse oximeters and invasive blood pressure transducers. Correspondingly, many successful lifesaving medical diagnoses and treatments, which were unthinkable before the inventions of the machines, are currently possible because of the valid measures of human body physiologic changes accurately determined by the modern equipment.

The Institute for Objective Measurement (2000) defined objective measurement as:

The repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured.

The definition implies achieving and preserving objective measurement of a variable, when using questionnaires or any other measuring tool, combining both mathematical and theoretical principles in translating observed raw data into abstract measures within acceptable error limits. In agreement, Thompson et al. (2005) argued that exemplary methods of measuring research variables are crucial to the quality and usefulness of evidence generated in a quantitative study. This argument suggests that only measures that align with the proven principles of measurement will be accepted as valid representatives of the amount of the variables examined notwithstanding the type of measuring tool used in generating the numbers.

Rethinking the validity of a questionnaire

To ensure a questionnaire is a valid measuring tool, basic or fundamentals of objective measurement must be rigorously inculcated, otherwise, the numbers allocated as measures will be grossly deficient and error-prone. However, most researchers are unaware of the consequences of failure to test measurement assumptions using advanced techniques reflecting Wilcox's (1998) concern that many researchers are using quantitative research methods which are not current, and lack skills required to utilise modern statistical methods. Hence, it is unscientific to assume any questionnaire measures the variable it purports to measure if the measures captured by the questionnaire are given the same treatment that some objective measures may be given, without demonstration of adequate evidence of validity of the questionnaire under the current measurement models.

Nurse researchers and clinicians, like their counterpart in medicine, education and social sciences, use questionnaires to gauge variables or traits of interest and by doing so generate numerical data which are eventually used for statistical analyses. Meanwhile, transferring and embedding the prerequisites of objective measurement to designing and analysing questionnaire data in nursing research has been mostly rudimentary and less rigorous (Hagquist, et al., 2009; Boone, et al., 2014; Blackman & Giles, 2015). For example, tallying and adding raw scores are common practices in analysing Likert-type items implying nurse researchers make no effort in translating the concrete answers into objective measures (The Program Committee of Institute for Objective Measurement, 2000). Additionally, most of the questionnaires used for data collection are not rigorously evaluated as required for variable measurement to be objectively conducted. Thompson et al. (2005) reiterated that researchers using questionnaires have ignored Wilcox's (1998) important warning that many powerful research outcomes have been lost or misunderstood because of failures to test essential measurement assumptions. Wright & Stone (1979) argued that measurement prerequisites are not mere scientific symbols but signposts of unacceptable error terms indicating a measuring tool has violated the objective standards. Interestingly, nurses are not novices to good measurement practice because many daily nursing interventions such as medication administration is firmly rooted in impeccable calculations. For example, if an adult patient is prescribed 1 gramme of paracetamol injection, nurses will follow through meticulously: calculating to the nearest exact amount of the prescribed medication and using an appropriate needle and syringe. Administering more or less of the amount of the medication may result in a range of outcomes from reporting a medication error incident to formally acknowledging a serious problem has occurred. Surprisingly, the high standard of measurement excellently demonstrated by nurses when administering medication to patients is trivialised among some researchers who use questionnaires for data collection due to the lack of a comprehensive approach to conducting and reporting the tested obligatory assumptions. By focusing more on the niceties of analysing data of low quality at the detriment of designing a good questionnaire for attaining a high-quality measure, most quantitative studies provide detailed reports of the data analysis process but an insufficient account of the psychometric properties of the measuring tool (Bond & Fox, 2007; Boone, 2016). Whereas applying the right statistical estimation is important, sophisticated analyses are not substitutes to excellent measurement practices. This

problem is intrinsic to the conventional approach to questionnaire validation frequently called the classical test theory or traditional validation methods (Hagquist, et al., 2009; Boone, et al., 2014; Sakib, et al., 2020).

Critical review: The limitations of conventional approach to questionnaire design

It is not that researchers using questionnaires are completely unaware of the need for exemplary validation techniques. Most questionnaires developed and used by nurse researchers have some forms of validation frequently referred to as “psychometric properties” in some nursing journals (Leung, et al., 2014; Melnyk, 2017). The evidence from systematic reviews of validity and reliability of questionnaires used in nursing research symbolises researchers’ emphasis on the quality of the measuring scales (Leung, et al., 2014). A good example of a scale validation and building process under the traditional or classical test theory is the 7- steps method that guided the design of the global health competency questionnaire (Stuhlmiller & Tolchard, 2018). Stuhlmiller & Tolchard (2018) outlined the processes of questionnaire development under CTT as: item generation; content adequacy assessment; questionnaire administration; construct validity; internal consistency assessment; concurrent validity; and use. The high point of this approach reiterates the need for a strong theoretical background and use of raw scores in estimating the construct under investigation.

However, this approach is inherently weak in generating good measures, because the raw scores used in the result computation are disordered and integral mathematical principles are assumed rather than statistically proven. Boone et al. (2014) noted that researchers in nursing and allied disciplines repeatedly cite Cronbach’s alpha value, factor analysis, convergent and/or concurrent validity as sufficient proofs of psychometric properties of a questionnaire. Meanwhile, psychometric validation has advanced beyond the severely weak traditional methods which are not rigorous enough to distil useful information of objective validity (Boone, 2016). Boone (2016) criticised the conventional concurrent and convergent validities which are based on comparisons between an old and new scale because the methods dismiss the possibility that the former scale in itself may lack validity when evaluated under the modern techniques. Moreover, of all the widely used classical tests, Cronbach’s alpha number is the most misapplied and misinterpreted psychometric parameter (Sijtsma, 2009). Sijtsma’s (2009, p.119) fourth (out of five) criticisms of alpha stated that alpha value cannot indicate dimensionality of a questionnaire because both

unidimensional and multidimensional questionnaires can possess very low or high alpha scores. The implication is that while alpha does not claim to assess dimensionality, the term 'internal consistency' that often accompanies most explanations of alpha value stifles the initiative to conduct a dimensionality test. However, for more than half a century, estimating the alpha value has become the most cited psychometric property of a questionnaire, thereby making Campbell's (1960) criticisms of alpha score more prophetic than conventional commentary. Following Cronbach's landmark article in *Psychometrika,* Campbell (1960) cautioned against researchers' targeting high alpha value at the detriment of the dimensionality of the variable under investigation; a warning ignored by many nurse researchers and social scientists, who repeatedly present the alpha value as the sole evidence of internal consistency or reliability or unidimensionality of a rating scale. But, having finally realised the limitations of some of the conventional psychometric evaluations, researchers' interest is becoming increasingly shifted towards exploring better methods of evaluating the psychometric properties of questionnaires without overstressing on the alpha value (Boone, 2016; Sakib, et al., 2020).

Another common problem with the past quantitative studies in nursing relates to the violation of the definition of measurement by adding up ordinal raw scores as a valid interval measure of a variable without due consideration for category functioning of the questionnaire (Bond & Fox, 2007). Category functioning test aims at finding out any disorderliness introduced to the category order. Bond & Fox (2007), building on the past work of Stevens (1946), defined measurement as consistent allocation of numbers to a construct or latent trait (such as belief or anxiety) underpinned by a predefined standard. When illustrated using a clinical thermometer, it means when a nurse measures body temperature, there is no arbitrariness: the reading is guided by a verifiable deflection in the level of liquid mercury in a glass thermometer. Applying the above definition of measurement to any questionnaire is to avoid arbitrariness in allocating numbers and adding up the figures as the final results of the scale. Meanwhile, a very good example of how measurement is disregarded by unproven calculations is recurring in the way past studies have treated questionnaires such as the evidence-based practice (EBP) belief scale. The EBP belief scale, developed by Melnyk et al. (2003), has 16 items with the ratings Strongly agree=5, Agree=4, Neutral= 3, Disagree=2, Strongly disagree=1. By computing the raw scores from respondents, one could easily produce a range of figures achievable from 16 to 80. But psychometricians consistently criticise adding up raw scores as valid measures either in a

polytomous Likert rating scale or dichotomous cognitive test design (Wright & Masters, 1982; Boone, et al., 2014; Blackman & Giles, 2015). The main problem here is that the five categories only tell one the category order, not the measure, whereas anyone doing interval scale computation with the raw scores from the EBP belief scale will erroneously come up with values that suggest 'Disagree=2' is half of 'Agree=4'. Until it is mathematically confirmed, one does not know the latent difference or gap among the categories and the numbers allocated to each category should not be added together for any serious statistical analysis. Thus, the starting point for developing a valid and reliable questionnaire is incorporating a methodology or technique that objectively demonstrates the dimensionality and category functioning of the questionnaire among other essential psychometric properties.

Advantages of Rasch techniques

In response to the problems, the Rasch techniques offer objective mechanisms that account for the dimensionality and category functioning (preservation of category order) as essential psychometric properties of a questionnaire (Linacre, 2004; Sakib, et al., 2020). The argument for Rasch analysis is demonstrated in the advantages offered in the stronger evidence of validity in the tested assumptions. While traditional data analysis treats participants' responses at face value by adding up raw scores, Rasch methods use probabilistic theory to investigate responses to each item or indicator that make up the questionnaire (Bond & Fox, 2007; Boone, et al., 2014). The Rasch equation states that in a test or survey, the probability of any level of endorsement, success or failure made by a person is a function of the person's level of ability (or knowledge or skill) and the level of agreeability (endorsement or difficulty) of the items (Wright & Master, 1982; Bond & Fox, 2007). The argument of this equation implies objective measurement by an item on a questionnaire is a function of interaction between the latent trait being measured and the level of agreeability or difficulty of the indicator. In other words, persons with lesser ability or knowledge face greater challenges in successfully endorsing difficult items while it is reasonable to expect that persons with higher ability or knowledge will find the same items easy to endorse. Additionally, the demand for objective measurement places an obligation on the questionnaire to show that all the indicators consistently target a single desired variable and not a contrast while simultaneously, disorderliness is not introduced to the predefined category order or options available to individual respondent. Rasch assumption tests confirming these two parameters of

objective measurement are called dimensionality and category functioning. Many other assumptions of objective measurement tested by the Rasch model emerged from the application of probability theory to analysing questionnaire or survey data.

Understanding unidimensionality of a questionnaire

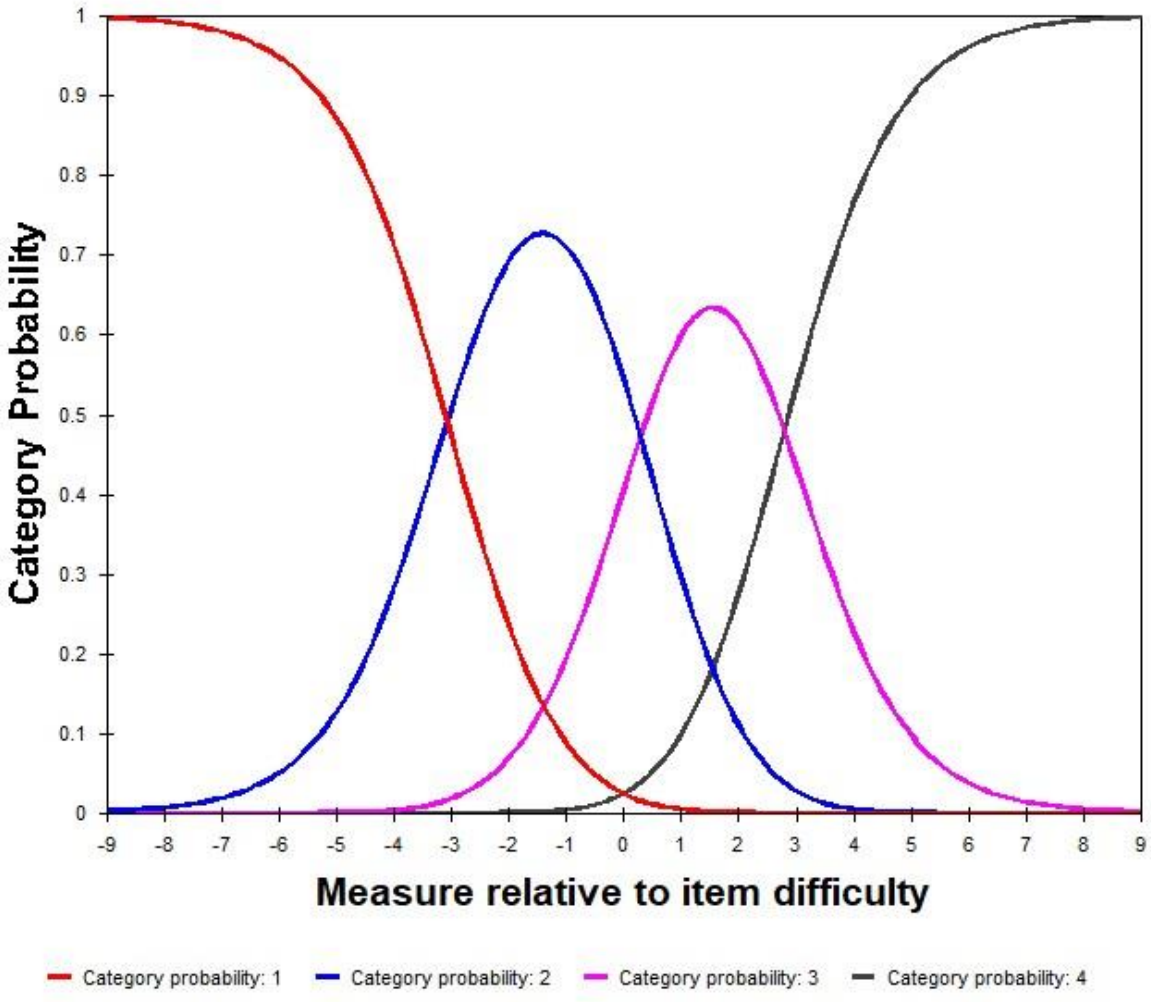
According to Segars (1997), a unidimensional questionnaire measures only one construct or variable and modern statistical testing allows investigation of this property to confirm or dispute the assumption. As most scales measure two or more latent variables per scale, confirmatory proof of unidimensionality of a tool occurs when all the indicators tap into the underlying variable. Mathematically, unidimensionality is displayed as parallel linear relationships of multiple items or indicators highly correlating with one underlying construct (Segars, 1997; Linacre, 2004). Measuring a research variable requires that multiple interrelated items are administered to the respondents with the intention the responses will typically display linear loading to the construct of interest. Arguing for dimensionality testing, Hagquist et al. (2009) explained that many latent constructs are closely interlinked and measuring a variable using multiple indicators become very difficult because of the interdependence and complex associations among the contrasting variables. The implication, according to Segars (1997), is that dimensionality testing confirms the composition of the research variable; however, researchers are obliged to clearly delineate the indicators based on evidence from relevant theories and literature. Two widely used statistical methods for assessing dimensionality are confirmatory factor analysis (CFA) and principal component analysis (PCAR) of residuals employed by the Rasch model (Sakib, et al., 2020). CFA is more strict and accurate than the PCA by restricting measurement to a predefined unit criterion (Joreskog & Sorbom, 1989). But Segars (1997) criticised CFA as possessing the likelihood of masking measurement errors due to its dual functions as measurement and structural equation model (SEM). Thus, Rasch proponents recommend conducting a PCAR which is a subsidiary of exploratory factor analysis (EFA) (Linacre, 2004; Boone, et al., 2014; Linacre, 2021). Unidimensionality of a scale is confirmed in PCA when the eigenvalue of the first unexplained residual is less than 2 and at least 50% variance of the factor is accounted for (Linacre, 2021). The most current computer tool for examining this property is table 23.0 under the output taskbar of the Winsteps Rasch analysis software developed by Mike Linacre (Linacre, 2021).

Understanding category functioning of a questionnaire

Another important psychometric property a questionnaire should display is the preservation of the category order implying participants using the options provided have minimised disorderliness to the predetermined category hierarchy. This quintessential requirement indicates that researchers cannot proceed with calculations based on the raw numerical data collected without, first, showing convincing proofs that the fundamental logic culminating with allocating higher numbers to higher category is followed by every respondent as specified in the questionnaire. According to Bond & Fox (2007), the evidence of good category functioning is underpinned by combining five parameters:

- Active category or observed category count: An active category in a measuring scale has high frequency, otherwise a category that is not well populated by participants is redundant and makes no meaningful contribution to the measurement.
- Normal distribution (Bell-shaped graph): Skewed categories, presenting with long tail and irregular distributions, are problematic to good measurement.
- Progressive measures: Average measures of each category should increase from the least to the highest indicating progression along with the average ability of the participants that endorsed the category.
- Category fit: Fit statistics display the degree to which data from a questionnaire comply with a Rasch model of measurement and value higher than 2 for a outfit mean square suggests injection of noise or misfitting of the category to the expected measurement model. Outfit value is sensitive to outliers, and the MNSQ value indicates the extent of noise or corruption injected into the measurement.
- Category threshold: Thresholds, also called Andrich thresholds or Rasch-Andrich threshold, is the probability level at which persons with higher ability moves up to endorse higher categories. Similar to category measures, thresholds of a good questionnaire will monotonically increase from the lowest category to the highest category in a good category function. The Winsteps computer software figure below is a typical example of a good category functioning of an item in a questionnaire (Figure 1).

Figure 1: A typical good functioning category of a test item



The vertical probability of the response axis shows the probability of endorsing a category ranging from 0 to 1 and the horizontal axis ranged from the least (-9) to maximum (+9) average measure achievable. From the output figure above, the first observation is the bell-shaped curves produced by the categories and the relatively equal gaps between each adjacent category indicating a normal distribution of data. Also, one can trace the point of intersections between adjacent categories to predict the category (Andrich) thresholds, which is the point on the graph at which participants with higher ability switched to a higher category. The category thresholds presented above increased monotonically from the least, category 1, to the highest, category 4, confirming a good category functioning. Advanced Rasch methods such as the

Partial Credit Model (PCM) simplify category thresholds in a polytomous rating scale (Bond & Fox, 2007). The PCM is a family of measurement theories that suggest intermediate efforts or responses introduced into a dichotomous scale represent a partial credit and should be scored accordingly, thus introducing category thresholds (Wright & Masters, 1982). Examples of Likert-type questionnaires assessed using the Rasch methods include children's empathic behaviours (Funk, et al., 2005), students' academic experiences (Curtis & Keeves, 2000), self-report EBP ability (Blackman & Giles, 2015), and most recently, fear of COVID-19 (Sakib, et al., 2020). Bond & Fox (2007, pg. 105) provided a comprehensive explanation of the measurement theories and introduction of item order to a Likert-type questionnaire. To clarify the technique, success or failure, in the Likert scale is phrased as 'fail to agree or fail to endorse or fail to approve' a category while introducing thresholds among the options or categories. Since the Rasch method is rooted in probability theory, a threshold implies the probability level beyond which failure to choose a category becomes the probability of endorsing a higher category (Bond & Fox, 2007). This means a dichotomous (Yes/No) test has a single threshold, a three-category scale will have two thresholds while a four-category Likert questionnaire will have three thresholds (Bond & Fox, 2007). Contrasting the traditional process where a rectangular or block analysis applies to all items by ignoring the lack of the equidistant in the category thresholds, Rasch-Andrich thresholds prove that thresholds for every item may vary significantly in a polytomous questionnaire (Linacre, 2004; Funk, et al., 2005). For instance, if an item from the EBP belief scale (I believe I can overcome barriers in implementing EBP) with response categories Strongly agree=5, Agree=4, Neutral=3, Disagree=2, Strongly disagree=1) is analysed using the classical methods, the gaps between each pair of adjacent categories will be displayed as one (1); strongly agree and agree is one (5-4=1) or agree and neutral (4-3=1) or neutral and disagree is one (3-2=1) This type of calculation is counterintuitive because the ordinal category or raw score is misinterpreted as interval or ratio where SA:A:N:D:SD is equivalent of 5:4:3:2:1. Higher numbers are allocated to the higher ordinal category (SA=5>A=4>N=3>D=2>SD=1) because participants who find any item extremely easy (to agree with) have a higher likelihood of choosing a higher category (SA=5), whereas those who find the same item extremely difficult to agree with will endorse the least category (SD=1). However, one must also realise that the jump from the lesser category to higher category (maybe SD to D or A to SA) is not the same for all participants and items, hence those who endorse SA=5 should be awarded only weighted

value of the category due to the variations in the category thresholds. Furthermore, since the only reason higher category is allocated a higher score is because of the higher level of ability or knowledge of the research variable, the thresholds along the categories must equally increase monotonically from lesser category (SD) to higher category (SA) in conformity with the predefined standard (Boone, et al., 2014). This is not achievable by some Likert questionnaires as thresholds within the categories may fail to increase proportionally (stagnate) or decrease, whereas category thresholds are completely ignored in classical analysis (Bond & Fox, 2007).

Limitations: Foremost, only two essential psychometric properties of a questionnaire are discussed here whereas there are many others such as item difficulty, local independence and differential item functioning. Secondly, this article suggested to researchers to consider evidence of dimensionality and category functioning of a questionnaire as prerequisites of objective measurement without empirically demonstrating the mathematical theories applied to any questionnaire used by nurses. Therefore, applying the psychometric validation techniques suggested here to many popular questionnaires used by nurses for clinical and research purposes will be very important for accurate and objective measurement.

Conclusion: Despite some of the advantages of modern psychometric validation techniques of questionnaires outlined here, the majority of nurse researchers designing and analysing questionnaires have not appropriated the benefits due to persisting widespread application of classical test theory (CTT) technique. In this paper, two major reasons are proposed for the continuation of traditional method or CTT practice even though a better mechanism or technique is available. First, the problem argued by Wilcox (1998) and Thompson et al. (2005) that many researchers engaging with quantitative design are not familiar with the application of many modern statistical techniques, although this concern is not limited to nurses and medical researchers alone. Secondly, some nurses designing questionnaires using CTT may not be aware of the criticisms of the traditional methods summarised below:

a. CTT fails to objectively test essential properties of the validity of a questionnaire, yet the untested assumptions are freely applied during data computations
b. CTT allows researchers to calculate variable measures using raw scores exported from the questionnaire, thus ignoring the fundamental mathematical requirement that indicators or items must be linearly related to one another before one can compute any point or score.

c. The misapplication and misinterpretation of some CTT assumptions such as the alpha value and concurrent validity as sufficient proofs of psychometric properties of a questionnaire also support the thinking that objective validity tests are not obligatory.

In contrast, it is argued that applying the techniques of objective measurement exemplified in Rasch methods will lead to producing high-quality questionnaires with better proofs of validity since integral mathematical principles are tested rather than assumed. This paper specifically encourages questionnaire developers to commit to rigorous examination of the dimensionality and category functioning tests of questionnaires. For a questionnaire to generate a valid measure of a research variable, all the items must be unidimensional by tapping into that particular variable only. To assess the dimensionality of a questionnaire, the Rasch theory suggests PCA of residuals of the items or indicators (Linacre, 2004; Boone, et al., 2014; Sakib, et al., 2020). Concurrently, a questionnaire category order must be evaluated to ensure the logical principle of allocating higher point or number to higher category is adhered to and not just assumed in the data collected. Assessing category functioning as an essential psychometric property involves evidence of sufficient category count, normal distribution of responses, good fitting of the category to the measurement model, good category thresholds and monotonically increasing average measures from the least to the highest category.

In conclusion, the usefulness of outcomes of a quantitative research study using questionnaires is strongly linked to inculcating objective and accurate measurement techniques. Traditional methods of assessing and analysing psychometric properties of a questionnaire are no longer tenable because the modern Rasch approach offers exemplary proofs of questionnaire validity rooted in objective measurement theories. Therefore, nurse researchers using questionnaire for clinical decisions and education purposes are encouraged to rethink the current approach to questionnaire design and analysis by asking themselves: Is this a good questionnaire? This is a question that can only be adequately answered satisfactorily by applying the fundamental principles of objective measurements demonstrated in Rasch theory.

References

Blackman, I. R., & Giles, R. (2015). Psychometric evaluation of a self-report evidence based practice tool using Rasch analysis. *Worldviews on Evidence-Based Nursing, 12*(5), 253-264.

Bond, T., & Fox, C. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences.* Mawah: NJ:Lawrence Erblaum & Associates.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, When, and How? *Research Methods*. doi:10.1187/cbe.16-04-0148

Boone, W., Staver, J., & Yale, M. (2014). *Rasch analysis in the human sciences.* Dordrecht: The Netherlands: Springer.

Colgrave, J., Stasa, H., & Fraser, J. (2020). Validity and reliability of the psychometric properties of a child abuse questionnaire. *Nurse Researcher*. doi:10.7748/nr.2020.e1677

Curtis, E., & Drennan, J. (2013). *Quantitative health research: Issues and methods.* Retrieved November 4, 2019, from http://www.ebookcentral.proquest.com

Curtis, E., Comiskey, C., & Dempsey, O. (2016). Importance and use of correlational research. *Nurse Researcher, 23*(6), 20-25. doi:10.7748/nr.2016.eI382

Funk, J. B., Fox, C. M., Chan, M., & Brouwer, J. (2005). The Development of the Children's Empathic Attitudes Questionnaire Using Classical and Rasch Analyses. *Journal of Clinical Child and Adolescent Psychology*.

Hagquist, C., Malin, B., & Gustavsson, J. P. (2009). Using the Rasch model in nursing research: A introduction and illustrative example. *International Journal of Nursing Studies, 46*, 380-393. Joreskog, K. G., & Sorbom, D. (1989). *Users; Reference Guide. Scientific Software.* Chicago: LISREL.

Leung, K., Tarvena, L., & Waters, D. (2014). Systematic review of instruments for measuring nurses' knowledge, skills and attitudes for evidence-based practice. *Journal of Nursing*, 2181-2195.

Linacre, J. (2004). Test validity and Rasch Measurement: Construct and content. *Journal of Applied Measurement, 3*(1), 970-971.

Linacre, J. M. (2021). *A user's guide to WINSTEPS MINISTEP Rasch-Model Computer Programs Program Manual 4.80.0.* Chicago: IL:Winsteps.

Melnyk, B. M. (2017). Models to guide implementation and sustainability of evidence-based practice: A call to action for further use and research. *Worldviews on Evidence-based Nursing, 14*(4), pp. 255-256.

Melnyk, B., & Fineout-Overholt, E. (2003, November 18). EBP Beliefs Scale.

Polit, D., & Beck, C. T. (2006). The contents validity index: Are you sure you know what's being reported? Critique and recommendations. *Research Nurse Health, 29*, 489-497.

Sakib, N., Bhuiyan, A. K., Hossain, S., AlMamun, F., Hosen, I., Adullah, A. H., . . . Mamun, M. A. (2020). Psychometric validation of the Bangla fear of COVID-19 Scale: Confirmatory factor analysis and Rasch analysis. *International Journal of Mental Health and Addiction*. doi:10.1007/s11469-02000289-x

Segars, A. H. (1997). Assessing the unidimensionality of measurement: A paradigm and illustration within the context of information systems research. *Omega International Journal of Management Science, 25*(1), 107-121.

Stuhlmiler, C., & Tolchard, B. (2018). Global health competency self-confidence scale: tool development and validation. *Global Health Science Practice, 6*(3), 528-537.

The Program Committe of Institute for Objective Measurement. (2000, December). *Definition of Objective Measurement*. Retrieved from www.rasch.org

Thompson, B., Diamond, K. E., Mcwilliam, R., Synder, P., & Synder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*(2), 181-194.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologists, 53*, 300-314.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design.* Chicago: MESA Press.

Dr Elizabeth Halcomb RN BN(Hons) PhD FACN

Editor, Nurse Researcher

Professor of Primary Health Care Nursing, University of Wollongong, Australia

Email: ehalcomb@uow.edu.au

RE: NR1842R1 - Your submission needs to be revised

Dear Editor,

We appreciate the painstaking efforts of your peer reviewers who read through our manuscript with the aim of recommending an improved last version for publication. We are equally very pleased that all the past issues raised in the first revision are now resolved without needing further corrections.

Nonetheless, the new editor (Reviewer 3) came up with some fresh criticisms on a section of our manuscript. We dissected through the reviewer's view and produced six points that the criticism probably highlighted. Correspondingly, we presented our responses to the feedback in red letters and offered rationale for our decisions on some parts of the criticisms which are not entirely applicable to the purpose of our article. Please find the copies of the article and our response attached to this email and author's submission.

In conclusion, we appreciate all the editors and your team for the clever work done so far. We are very hopeful that you get this manuscript published as all the reviewers have attested to its quality, robustness, and scholarship. Thank you in anticipation of your favourable consideration and timely response to this email.

Yours sincerely,

Odunayo Kolawole Omolade.

Reviewer's Criticisms	Responses
The authors have addressed reviewers' comments and have made sufficient revisions. However, the content under the subheading "Critical review: The limitations of conventional approach to questionnaire design" requires a revision. A significant gap is lack of details or acknowledgement of the key fundamental process that are involved in the initial development of a questionnaire that do not involve the use of Rasch techniques.	Thank you for this comment. We had previously referred to the classical approach to questionnaire design laid out by Stuhlmiller & Tolchard (2018) and have expanded on this to include a commentary on the extent of the rigour in which these steps may be followed. Hopefully in doing so we have strengthened our rationale for the use of Rasch methods whilst acknowledging that a well-constructed tool using traditional methods is not without value. Please note that we also mentioned some systematic literature review of psychometric properties of questionnaire, and we summarised the main contributions of this method as reiterating theoretical background of the questionnaire and use of raw scores for variable estimation. We believe readers can familiarise themselves with the methods by reading the suggested articles and similar ones; negating the need for presentation of detail of classical test theory, which may distract from the main focus of our study.

Explanations for disagreement

Reviewers' criticisms	Responses
The methods described in the manuscript are only applicable after a questionnaire has been developed. There are important considerations and well-established processes that must be followed in the initial development of a questionnaire before any validity can be established.	The reviewer argued Rasch technique is only relevant to a questionnaire already "developed". However, we argue that a questionnaire is not as developed for generating valid measurement until proofs of objective parameters ratified in the questionnaire are displayed. The second part of the reviewer's argument is concerned with, again, what the reviewer termed "the initial" process, a repeat of the first criticism. We believe that the revisions made to the article supported by the explanations we provided have addressed this criticism. For emphasis, the use of Rasch techniques does not, of course, eliminate the need for all of the steps laid out by Stuhlmiller and Tolchard to be conducted to a certain level. In particular, Rasch analysis does not preclude the need for a suitable set of constituent questions to be derived using a rigorous iterative process such as the Delphi process and other content statistical analyses. But where Rasch methods are applied, it could

	be argued that the need for CTT statistics is reduced.
There are references to "designing" a questionnaire without providing sufficient context.	Thank you for this comment on the context of our article. In the beginning, we narrowed down the context of designing a questionnaire by focusing solely on questionnaire measurement in nursing. We argue that the purpose of a questionnaire, used in quantitative nursing research, is to measure a variable or variables of interest. All our examples and illustrations are consistent with the nursing profession. Hence, we disagree that there is not sufficient context to this argument.
Illustrations involving medication administration errors and medication error incidents are also provided without acknowledging the important steps in the development (pharmacodynamics and pharmacokinetics) of the drug formulation.	Thank you for this comment. We aimed to limit our examples to simple drug calculations by nurses only. The central idea is that we suggested to nurse-researchers developing questionnaires to imbibe the rigours and flawlessness demonstrated by their clinician's colleague during drug calculations. We would suggest that it is not necessary, and less relevant, to introduce controversial issues such as drug metabolism and excretion: noting in this specific case that when a nurse wants to calculate the doses of prescribed medication, no known drug calculation formula is used at the bedside incorporates pharmacodynamics and pharmacokinetics of the drug.
Medication administration is the last step in the process, the same for the use of Rasch techniques- they are only relevant after the initial development of the questionnaire in question.	Thank you for this comment. The argument that medication administration is the last step in the process is inconsistent with the nursing process theory which specifies the evaluation of the therapeutic outcome (or side effect) of a drug administered as the last step of nursing actions. Hence, applied to the current paper, Rasch techniques represent assessing the psychometric properties of the questionnaire, rather than the last step in the process which is the revision of the questionnaire items.
In addition, the authors must acknowledge that there is no silver bullet when it comes to establishing the reliability and validity of a questionnaire is an iterative process and further psychometric validation is always warranted in different settings, contexts, cultures and languages.	Again, we thank the reviewer for this feedback. Under the subheading "limitations", we presented some relevant criticisms of our article. As researchers, we believe in scientific advances for the same reason we encourage readers of this article to advance their questionnaire validation techniques. Our position is that the Rasch techniques are more advanced than the classical

	test theory yet both theories can be applied to complement each other. Finally, setting a gold standard for psychometric assessment of questionnaires is not our prerogative, rather the goal of our article is to raise the standard by advocating for the application of the most current objective techniques shown in Rasch methods of questionnaire validation.