NIRJ
NKUMBA INTERNATIONAL RESEARCH JOURNAL

Nkumba International Research Journal (NIRJ)

Volume 2 Issue1, May 2022

Advancing Psychometrics in Uganda's Institutions of Higher Learning with Emphasis on Structural Equation Modeling (SEM)

Norman David Nsereko[1], Stephen Palmer[2], Veronika Basa[3]

Corresponding author: Assoc Prof. Norman D Nsereko
Nkumba University, P. O. Box 237 Entebbe, Uganda,
Phone: +256 703 335 948 Email: nnsereko@nkumbauniversity.ac.ug

[1] Assoc Prof. Norman D Nsereko (PhD) is an Associate Professor of Counselling Psychology in the School of Social Sciences at Nkumba University. He is a researcher, publisher, and developer of the multi-dimensional UNIVERSITY STUDENTS EVALUATION OF PSYCHOSOCIAL PROBLEMS (USEPP) scale. http://orcid.org/0000-0001-9461-6057

[2] Prof Stephen Palmer (PhD) is a Professor of Practice at the University of Wales Trinity Saint David, and Director of the Centre for Stress Management, London. He has written and edited 60 books and has published over 225 articles. https://orcid.org/0000-0002-0108-6999

[3] Veronika Basa is an independent researcher, author, course designer, and developer of the first nationally accredited courses in supervision, within the Australian Qualification Framework (AQF), in Australia. http://orcid.org/0000-0002-6908-4930

Abstract

In Uganda and other countries, many academic research projects are carried out in various disciplines e.g., health, social, behavioral sciences, etc., especially for an academic award. Data gathering to measure the attributes of the study is sometimes carried out by applying instruments with incomplete scale development although a large amount of technical literature on scale theory and development exists.

Experts in psychometrics advise researchers to use existing and validated instruments suited for an attribute or construct for measurement in one's research project to obtain reliable and valid data or undertake the comprehensive and scientifically rigorous process of developing an instrument to use for one's research project. This observational and documentary study has addressed the knowledge and practical gaps in scale development while carrying out scientific inquiry informed by the fundamental concepts in psychometrics theory which is often not a part of graduate training in Uganda's institutions of higher learning.

Keywords: *Psychometrics, Scale development, and evaluation, Item Response Theory, Structure Equational Modeling, Model Specification, Model Identification, Model Estimation, Model Testing.*

Introduction

While the primary aim of psychometrics is to develop psychological instruments to improve psychological science based on different theories and approaches, it also encompasses the mathematical, statistical,

and professional methods that address how tests are constructed and used, and indeed, how they are evaluated in different disciplines e.g., health sciences, business administration, social sciences and education (Anunciação, 2018; Buchanan & Finch, 2005).

With the advancements in computer and software technologies, psychometrics has developed new methods of statistical analysis or the refinement of older techniques prompting the growth in the use of statistical and psychometric methods in psychological, social, and educational research (Anunciação, 2018). Free technologies have become available that can assist researchers who lack resources. For example, the GPower software can be used for undertaking different types of power analysis (Faul et al., 2007; 2009).

Many academic research projects are carried out in various disciplines e.g., health, social, behavioral sciences, etc., especially for an academic award. Data gathering to measure the attributes of the study is sometimes carried out by applying instruments with incomplete scale development. Understanding the approaches/paradigms e.g., the Item Response Theory (IRT) used in evaluating the quality of tests under psychometrics will enhance the abilities of researchers to construct psychometrically valid tests in their work.

Psychometrics in the Intervention of Human Needs

Psychometrics has mainly dealt with the understanding of human constructs. These include intelligence, interests, abilities, aptitude, personality, behaviors, and their antecedents/consequences namely; cognitive, sensory, perceptual, or motor functions. Scales are therefore developed to measure the constructs of interest and applied in the day-to-day life experiences. The scales developed are referred to as psychological tests because they are standardized procedures for sampling behavior and describing it using scores or categories. Tests may be predictive of some non-test behavior of interest or they may be norm-referenced, describing behavior in terms of norms (Jones & Thissen, 2007).

Psychometrics covers three broad areas. (1) Psychological scaling: this deals with a set of techniques for the assignment of quantitative values to objects or events using data obtained from human judgment. (2) Factor analysis: this deals with methods and procedures used to explain the observed covariation among a set of variables in relation to the underlying or latent (unobserved) variables. The emergency of factor analysis made other aptitude and personality tests possible. (3) Psychological measurement: this combines with psychological scaling and factor analysis to produce a new test theory characterized by item response theory (Jones & Thissen, 2007).

Although psychometrics has generally been viewed as a value-free discipline, it is value-laden due to the four values, namely: that individual differences are quantitative and not qualitative, that measurement is objective in a specific sense, that test items are fair, and that a model's usefulness is more important than its truth (Wijsen, Borsboom, & Alexandrova, 2021).

Techniques required for scale development and evaluation

Thompson, Loesch, and Seraphine (2003) observed that for an assessment instrument to be both credible and usable, it must be grounded in a framework that enjoys widespread and substantive endorsement. Instrument development is based on different theories and approaches. The primary measurement theories in psychometrics used by researchers to construct psychological assessment instruments are Item Response Theory (IRT) and Classical Test Theory (CTT) (Anunciação, 2018).

These two distinct models/approaches are used in evaluating the quality of tests. Both IRT and CTT deal with broad concepts, including reliability and validity.

The Classical Test Theory is used mostly to evaluate structural validity whereas the Generalizability Theory is more concerned with how instruments are applied to measurement activities when applied with a two-step sampling procedure (Jones & Thissen, 2007; Bjorner et al., 2005; Fischer & Molenaar, 1995). Currently, IRT and CTT methods are seen as complementary and are frequently used to assess

the test validity and respond to other research questions (Anunciação, 2018). Our broad interest and emphasis will look at IRT.

Item Response Theory

Item Response Theory represents a collection of statistical models and methods used in psychological measurements (test theory) mainly for two broad purposes in the measure of health outcomes: item analysis and scale scoring. Basically, IRT addresses the evaluation and improvement of the basic psychometric properties of items and tests (Furr & Bacharach, 2008). IRT models handle unidimensional data as well as multidimensional data, binary and polytomous response data, and ordered as well as unordered response data (Jones & Thissen, 2007; Van der Linden & Hambleton, 1997).

IRT focuses on the statistical analysis of test data on the person's responses to survey items, and his or her standing on the construct being measured by the scale rather than on the total summed test score. It is usually applied when examining very large samples with the goal of generalizing the results to a broad population. It is a much more flexible and informative theory since it provides item statistics that are population-independent, over other theories used in factor analysis i.e., The Classical Test Theory. Specifically, most IRT models and methods predict that one or more unobserved (latent) traits and item parameters underlie the responses to test items whereby the variation among individuals on those latent variables explains the observed covariation among item responses (Jones & Thissen, 2007; Bjorner, Kosinski, & Ware, 2005; Fischer & Molenaar, 1995). An approach to IRT in scales development gives a good psychometric grounding of the scale (Reijneveld et al., 2003).

Instrument development in the Ugandan context

Instrument development to reflect the contextual variables in the Ugandan situation and its application is still at its lowest. There are few instruments reported in Uganda that have been locally developed and validated for clinical or research purposes in mental health. Ovuga (2005) developed the Response Inventory for Stressful Life Events (RISLE) and revised it from 100 items to 36. It measures depression and suicidal ideation/ attempts. The RISLE omitted a confirmatory factor analysis procedure to determine whether the hypothesized factor structure provided a good fit to the data, in other words, whether or not a relationship between the observed variables and their unobserved or underlying latent constructs existed.

A recently locally-developed and validated instrument is code-named the University Students Evaluation of Psychosocial Problems (USEPP). It is a multidimensional, self-administered psychological instrument measuring psychosocial problems among university students. It can detect university students with and those without psychosocial problems and it can predict psychological distress (Nsereko, Musisi, & Holtzman, 2014).

Other Ugandan researchers have attempted adaptation and modification methods of already developed instruments to make them contextually and culturally relevant to Ugandan settings mainly for research purposes (Ovuga et al., 2006; Nakimuli-Mpungu, Musisi, Katabira, Nachega, & Brass, 2011; Nakigudde, Musisi, Ehnvall, Airaksinen, & Agren, 2009; Abbo et al., 2009). However, these will still lack the cultural and contextual specificity for the Ugandan situation.

Academic research projects for an academic award

A research project is an academic imperative for graduation for students at bachelor's, master's, and PhD levels in the Ugandan education system. It is presumed that the student researcher will investigate the variables of study in a systematic manner and analyze the data obtained in an accepted statistical procedure. Students undertake such a project involving gathering empirical data in an effort to build new models and paradigms.

The widest approach used to collect data in survey designs is the use of questionnaires and other quantitative tools when one needs to accurately measure the variables of interest (Stone, 1978). However, where these efforts have been applied to use instruments for assessing clients or for research purposes it has been mainly done by reference to Eurocentric measures that are on the market.

Alternatively, as experience has shown, student researchers attempt to develop an instrument to measure the manifest variable observed in the real world, or something else measured as part of a theory. It is important that measures on these survey instruments adequately represent the constructs under examination (Hinkin, 1998; Barrett, 1972) to come up with a well-constructed, valid and reliable test that has been subjected to a comprehensive and scientifically rigorous process of development that stands the test of time. Schoenfeldt (1984) underscores the importance of sound measurement as follows: "The construction of the measuring devices is perhaps the most important segment of any study. Many well-conceived research studies have never seen the light of day because of flawed measures" (p. 78).

When one takes an interest to read some of the research reports submitted for graduation, one discovers that the instruments that are normally developed by student researchers are not comprehensive. They normally end at the initial stage of item development involving coming up with the initial set of questions for an eventual scale, covering the following: "(1) identification of the domain(s) and item generation, and (2) consideration of content validity and the calculation of Internal consistency reliability" (Boateng, et al., 2018, p2).

This incomplete scale development process is quite a commonplace occurrence as there are several incomplete scales used to measure mental, physical, and behavioral attributes (Bai, Peng & Fly, 2008; Hirani, Karmaliani, Christie, & Rafique, 2013), although a large amount of technical literature on scale theory and development exists (DeVellis, 2012; Ajzen, 1985). Experts in psychometrics advise researchers in this field to use existing and validated instruments suited for an attribute or construct for measurement in one's research project to obtain reliable and valid data or undertake the comprehensive and scientifically rigorous process of developing an instrument to use for one's research project. Anything less could undermine the credibility of the research findings (Worthington & Whittaker, 2006; Hooper et al., 2012).

Several factors have been forwarded as to why the advice of the test developer experts is not adhered to especially among researchers in our local universities. Among these include: Existing and developed instruments that would address the researchers' constructs, attributes for the study are too expensive to purchase and they may not be multiculturally appropriate to address local issues of interest. And the most compelling reason rests with the non-existence of training programs that address instrument development in our universities that cover a comprehensive and scientifically rigorous process of developing an instrument for research purposes. Other researchers point to the copyright imperative that discourages those who might be in possession of validated instruments. Still, others point out that some manuals accompanying the instruments of interest are rather hard to understand. Moreover, others argue that developing an instrument is very rigorous and expensive, time-consuming given the time allotted to complete the research project.

These anecdotal observations do rhyme with researched data. For instance, Boateng et al., (2018) noted that although using an existing questionnaire will optimize time and resources, a questionnaire that measures the construct of interest may not be readily available, or the published questionnaire is not available in the language required for the targeted respondents. Many universities do not include training on scale development. Others do not have a well-established framework to guide researchers through the various stages of scale development (Price & Mueller, 1986). There is poor reporting of newly developed measures that are used for research that threaten the reliability of data (Hinkin,1995).

Senyonyi et al., (2012) observed that the scales that are already developed specifically for assessing student needs are, in most cases, outdated, expensive, or inaccessible, and are often not contextualized or validated in Ugandan settings. They are inadequate to capture the specific contextual variables of the area of study (Ahia & Bradley, 1984). Sometimes, when no effort is taken to validate them, it complicates issues of reliability and validity.

The common practice in instrument development

In Uganda, many researchers quite often attempt to develop their own instruments. They do so because of a failure to establish an appropriately developed and validated instrument for their studies due to several reasons as cited above. A number of completed undergraduate, master's and even PhD studies fall under such categories (personal observation of some studies which have already gained credit). The

common characteristic of these instruments is that the instrument developers usually accomplish the first step of the model specification where a pool of items is generated to tap the construct to arrive at a set of items that should represent the construct of interest. They miss out on important imperatives of psychometric principles on instrument development like applying theories underpinning the process and other phases to creating a comprehensive process of developing an instrument for their research project. This creates a big hurdle in designing a questionnaire that is psychometrically sound and is efficient and effective for use in research and clinical settings.

Based on the IRT frameworks, we continue to observe that they fail to accomplish factor-analytic, data-reduction techniques, and scale evaluation techniques that would yield a stable set of underlying factors that accurately reflect the construct (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018).

These instruments also do not report construct and predictive validity. The reliability measures are arbitrarily derived by reference to test and retest methods which cannot be sustained because the items of the scales and the latent variables were not psychometrically derived in the first place. Even if at this stage many tests indicate internal consistency reliability, it is not a sufficient condition for construct validity (APA, 1995).

If the right psychometric properties of the instrument are not established first, that creates methodological issues in research studies (Worthington & Whittaker, 2006). It also falls short of the norm and central objective of psychometrics which aims at ensuring that psychometric qualities remain up to date (Osborne, 2010).

Consequently, such a scenario throws the quality these research projects and findings in Uganda in doubt as well as the decisions and awards that may be based on such efforts, including academic credentials and policy decisions.

A serious drawback in promoting sound instrument development practices in the Ugandan research sector is that universities do not include research courses on psychometrics, factor analysis (exploratory, confirmatory), and Structural Equation Modeling (SEM) analysis techniques, at both the master's and doctoral degree levels

Scale development informed by Structural Equation Modeling (SEM)

This section intends to demonstrate how a researcher intending to develop an instrument for research or clinical purposes applies an advanced statistical multivariate analysis technique referred to as Structural Equation Modeling (SEM), a derivative of Item Response Theory (IRT). SEM is more of a confirmatory technique, but it can also be used for exploratory purposes (Schreiber, et al., 2006)

SEM models combine the path and factor analytic models to explain relations among a large set of observed variables using a small number of latent/unobserved variables, allowing a researcher to confirm the factor structure of a newly developed instrument (Crockett, 2012; Anunciação, 2018)—more specifically, "SEM tests models that specify how groups of variables define a construct, as well as the relationships among constructs" (Crockett, 2012, P.2).

When performing SEM, large samples are essential (asymptotic) to provide stable parameter estimates (Bentler, 1990; Savalei & Rhemtulla, 2017). There are different opinions on the exact sample size. Worthington and Whittaker (2006) recommend the Bentler and Chou's (1987) guideline of at least the 5:1 ratio of participants to a number of parameters, with the ratio of 10:1 being optimal. Other scale developers recommend a sample size of 150 respondents in exploratory factor analysis and a minimum sample size of 200 for confirmatory factor analysis (Hinkin, 1998).

Steps in Performing SEM Analysis

There are five applied steps for conducting SEM analysis: Model Specification, Model Identification, Model Estimation, Model Testing, Model Modification.

1. Model Specification

The first step of SEM analysis is model specification. Model specification occurs prior to data collection and analysis. It is aimed at defining the phenomena of interest e.g., cognitive, behavioral, and psychosocial problems, etc., clearly and concretely, using both existing theory and research to provide a sound conceptual foundation (Worthington & Whittaker, 2006, p. 813).

The phenomena of interest or the construct help to guide item generation so that factor-analytic, data-reduction techniques yield a stable set of underlying factors that accurately reflect the construct and the validity of its content (Worthington & Whittaker, 2006).

In scale development literature, model specification is referred to as the logical-content or rational/theoretical approach. The logical approach applies the scale developer's judgments to identify or construct items that represent the characteristic being measured. The logical approach generates items through literature review and assessment of existing scales and indicators of a domain under consideration. The rational/theoretical approach both begins and ends with informed judgments based on theory. It generates scale items from individual interviews, direct observations, and focus groups (Boateng, et al., 2018; Hinkin, 1998).

Once the items for the model have been identified, expert opinion is sought to evaluate the items for analysis of content validity (e.g., the extent to which a set of items reflects the content domain) readability and wording, the adequacy of the items in terms of contextual relevancy of the items, language simplicity, length, format, and inclusiveness. At this stage, experts may offer suggestions e.g., for adding new items or deleting some.

When the model is created, the most likely explanations for relationships included in the model and a rationale for the overall specification of the model should be given (Crockett, 2012). The items developed for the model are typically self-report in nature and tend to utilize Likert-type agreement response scales. These generally have three to seven choices ranging from Strongly Disagree to Strongly Agree.

Although the items constituting a particular assessment can be grouped under an overall construct (e.g., psychosocial problems), it is common for subsets of items to describe specific facets within the larger construct (e.g., emotional problems, traumatic problems, academic problems, antisocial behavior (Holtzman & Vezzu, 2011). Finally, the items are administered to a development sample, after which they are evaluated and the scale length is determined (Worthington & Whittaker, 2006).

However, stopping at the logical, rational/theoretical approach is no longer a popular method in scale development for any kind of purpose (Crockett, 2012; Worthington & Whittaker, 2006; Hooper et al., 2012).

What is widely advanced in instrument development is the empirical procedure that employs factor analysis (exploratory, confirmatory) and Structural Equation Modeling (SEM) analysis techniques (Crockett, 2012, Schreiber, Nora, Stage, Barlow & King, 2006).

2. Model Identification, Model Estimation, Model Testing, Model Modification

Model identification, Model estimation, Model testing, and Model Modification are the second, third, fourth, and fifth steps respectively carried out in SEM analysis. They constitute the empirical approach to scale development. They use factor analysis to identify a set of factors representing underlying latent constructs from a larger number of observed variables of items on a survey. Factor analytic techniques, properly employed, help to determine (a) predictive utility for a criterion group (e.g., psychosocial problems) or (b) homogenous item groupings i.e., whether groupings of the observed variables/items on a survey demonstrate the psychometric properties necessary to assert they reliably and validly measure one or more intended constructs (Worthington & Whittaker, 2006).

Model Identification

Model identification occurs before estimating model parameters (i.e., relationships among variables in the model). It is intended to identify the theorized model by looking at the possible unique solution generated for each parameter in the model and whether it is dependent on the designation of model parameters.

The identified model should have the following characteristics:(a) there are two or more latent variables, each with at least three indicators that load on it, the errors of these indicators are not correlated, and each indicator loads on only one factor, or (b) there are two or more latent variables, but there is a latent variable on which only two indicators load, the errors of the indicators are not correlated, each indicator loads on only one factor, and the variances or covariances between factors is zero (Crockett, 2012; Worthington & Whittaker, 2006). SEM statistical software programs allow a researcher to determine whether the model is identified or not.

Two complementary statistical analyses namely Exploratory Factor Analysis (EFA) for model identification and estimation and Confirmatory Factor Analysis (CFA) for model testing are applied sequentially. EFA explores the factor structure of the responses to some set of survey items, while CFA is used to confirm whether specified groupings of items properly measure the theorized constructs of interest (Crockett, 2012; Worthington & Whittaker, 2006). (These two analysis techniques are elaborated below).

3. Model Estimation

Model estimation follows model identification. It involves estimating the parameters of the theoretical model in such a way that the theoretical parameter values yield a covariance matrix as close as possible to the observed covariance matrix S (i.e., the matrix derived from the sample data) (Crockett, 2012).

4. Model Testing

Model testing involves the analysis of two conceptually distinct models: the measurement model and the structural model. The process determines whether the hypothesized structure provides a good fit to the data, or in other words, whether a relationship between the observed variables and their underlying latent, or unobserved constructs exists (Child, 1990). It also verifies that all items are properly aligned with the correct facets within the general construct being measured. CFA is conducted on the measurement model to effect model testing (Crockett, 2012). Raykov and Marcoulides (2000), Stage et al. (2004) as cited in Crockett (2021) advise that a path diagram or visual representation of the hypothesized relationships among variables are part and parcel of the SEM procedures and must be shown

5. Model Modification

This step may be carried out at the EFA stage when there is a need to trim or add new parameters in an attempt to improve the theoretical model's fit to the data. The process is restricted to the initial data worked on in the EFA (Crockett, 2012).

Exploratory Factor Analysis

Researchers perform EFA followed by CFA when constructing new scales. The survey is first administered to a representative sample and the data are subjected to an EFA, and the factor structure uncovered by the EFA is subjected to a CFA using data collected from a new sample.

The initial step of data analysis involves screening item responses for normality because the extraction methods like Maximum Likelihood Estimation in factor analysis assume normality of the item responses. Skewness and kurtosis for the item responses are evaluated to examine normality (Worthington & Whittaker, 2006).

Sample Size

There are differing opinions in the literature about sample size in the development of an instrument. Some studies have made different suggestions regarding sample sizes for carrying out Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) when developing an instrument. For instance, Hatcher (1994) advocated the minimal sample size to be larger than 100 participants or 5 times the number of variables being analyzed. Reise, Waller, and Comrey (2000); Thompson (2004) considered sample sizes less than 100 or with fewer than 3:1 participant-to-item ratios as generally inadequate.

Worthington and Whittaker (2006) who have spelled out comprehensive guidelines for scale development recommended obtaining a sample of about 300 respondents and more in the initial use of EFA. Following the outcome of the EFA, they also suggested that data collection for use in CFA should at least be 200 respondents.

In general, there is some agreement on large samples in scale development because scale variance attributable to specific participants is likely to be canceled by random effects as sample size increases (Tabachnick & Fidell, 2001). Larger sample sizes are likely to result in more stable correlations among variables and will result in greater replicability of EFA outcomes (Worthington & Whittaker, 2006).

Factorability of the Scale

Following item response screening for normality, data are then verified for factorability (correlation matrix) before factor extraction. Researchers can use Bartlett's (1950) test of sphericity to estimate the probability that correlations in a matrix are 0. The disadvantage of this approach lies in its being highly susceptible to the influence of sample size and likely to be significant for large samples with relatively small correlations (Tabachnick & Fidell, 2001). Or they can use the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy to determine the appropriateness of factor analysis.

KMO indicates the extent to which a correlation matrix actually contains factors or simply chance correlations between a small subset of variables. Values of .60 and higher are mandatory for good factor analysis while values greater than .60 are considered to be adequate and greater than .80 are considered to be high. (Tabachnick & Fidell, 2007).

Extraction Methods

After establishing the factorability of the scale, the latent variable underlying the scale is determined through a variety of types of extraction methods in EFA. These include principal components, image factoring, unweighted least squares, generalized least squares, alfa factoring, principal axis factoring, and maximum likelihood. Before any technique is adopted researchers need to understand the distinct purposes of each technique. For instance, principal-components analysis reduces the number of items while retaining as much of the original item variance as possible. While principal axis factoring and maximum likelihood aim at understanding the latent factors or constructs that account for the shared variance among items. They are more closely aligned with the development of new scales since they assume multivariate normality and lack of skew (Worthington & Whittaker, 2006).

Rotation Methods

Factor extraction is followed by the use of rotation methods. These include Varimax, Direct Oblimin, Quartimax, Equamax, and Promax. Factor extraction aims to understand and interpret the latent factors or constructs that account for the shared variance among the items, and purposes of subsequent use in regression or other CFA prediction techniques. When the factors are expected to be correlated, the Oblique Rotation method e.g. Direct Oblimin is preferred while for uncorrelated factors the orthogonal rotation method e.g. varimax is used (Worthington & Whittaker, 2006).

Determining the Number of Factors to Retain

There are several guidelines followed to determine the number of factors to retain for the scale. The commonly used methods were suggested by Kaiser (1958) and Cattell (1966) based on eigenvalues.

These approaches namely the traditional eigenvalue cut-off of 1 and the scree plot (the relative values of eigenvalues) may help to determine the importance of a factor and indicate the amount of variance in the entire set of items accounted for by a given factor. Eigenvalues that are less than 1.0 reflect potentially unstable factors for retention and they are naturally excluded for further analysis (Worthington & Whittaker, 2006; Tabachnick & Fidell, 2007).

Factors can also be decided based on the criteria for a simple structure. These include factors that have three items and above with conceptual interpretability. The items that have high loadings on each factor which are greater than ≥ .32 and do not double-load (a.k.a., cross-load) onto any other factors at the ≥ .32 measure. Researchers with their discretion can retain a factor with only two items if the items are highly correlated (i.e., r > .70) and relatively uncorrelated with other variables. But above all a specific factor structure to be retained, must be adequately reproduced during EFA (Worthington & Whittaker, 2006). Items that do not fulfill these criteria are otherwise dropped because they may be considered insufficient indicators of the factors produced in the EFA and the criteria used are reported. However, when multiple iterations of EFAs are run, some dropped items may be reintroduced in the analysis at the researcher's discretion.

Confirmatory Factor Analysis

CFA is a special case of the structural equation model (SEM), also known as the covariance structure (McDonald, 1978) or the linear structural relationship (LISREL) model (Jöreskog & Sörbom, 2006). SEM typically consists of two models: a measurement model linking a set of observed variables to a usually smaller set of latent variables and a structural model linking the latent variables through a series of recursive and non-recursive relationships. The casual relationships are aimed at determining whether the specified model is identified (Schreiber, et al., 2006). An unidentified model is one for which it is impossible to come up with unique parameter estimates. CFA corresponds to the measurement model of SEM and it is estimated using SEM software. The most common SEM software used include Amos, LISREL, and Mplus.

Once a factor solution has been established, it is important to perform confirmatory factor analysis (CFA). The CFA process determines whether the hypothesized structure provides a good fit to the data, or in other words, whether a relationship between the observed variables and their underlying latent, or unobserved, constructs exists (Child, 1990).

It involves the analysis of two conceptually distinct models: the measurement model and the structural model. New data is used from another sample or the data from the same sample can be randomly split for each analysis. This is done to determine whether the proposed measurement model obtained in the initial EFA holds, ensuring that the chosen observed indicators, items, and the latent constructs identified find a good fit for the model. The CFA would also verify that all items were properly aligned with the correct facets within the general construct being measured and thus help support the factor structure reliability and the validity of the scale (Crockett, 2012).

Model Fit Indices

Model fit indices are used while running CFA to indicate how the model that best represents the data reflects the underlying theory. Several goodness-of-fit indicators have been developed and they continue to evolve. When the model-fit indices are suitable, after considering their application to different samples sizes, types of data, and ranges of acceptable scores then the theoretical model is considered to be supported by the sample data. Model-fit indices that fail the various recommendations indicate that the sample data do not support the hypothesized model, requiring the re-specification of the theoretical model (Crockett, 2012; Schreiber, et, al., 2006).

The use of model fit indices applied in CFA and the cutoff values when fitting structural equation models is still a matter of debate (Worthington & Whittaker, 2006). A model fit is by no means agreed upon since statisticians are always seeking and developing new and improved indices that reflect some facet of model fit previously not accounted for (Hooper et al., 2012). Some experts in the field look at fit indices with the like hood of the incorrect rejection of an acceptable model or retaining one which has poorly fitting parts (Marsh et al., 2004).

Others think that overreliance on the model fits may compromise the original, theory-testing purpose of structural equation modeling (Hooper et al., 2012). Others have denounced the use of fit indices altogether (Barrett, 2007). Nevertheless, studies on scale development include fit values in their reports, and editors, reviewers, and consumers look for recommended values when assessing a sound scale which indicate the importance of model fit indices (Schreiber et. al., 2006).

Model fit indices are classified under three distinctive types of fitting functions: absolute, incremental, and parsimonious.

Absolute Fit Indices

Absolute fit indices determine how well the structural model reproduces the sample covariance matrix sample and demonstrates which proposed model has the most superior fit. These goodness of fit indices are considered the most fundamental indicators of how well the proposed theory fits the data (Schreiber et al., 2006). Examples include:

1. **Chi-Squared test**: The model chi-square with corresponding degrees of freedom and level of statistical significance. A chi-square value close to zero and a chi-square p-value greater than 0.05 indicate that there is little difference between the expected and observed covariance matrices, which is one indicator of good fit (Hooper et al., 2012). When the chi-square test fails to give a good fit, Wheaton, Muthen, Alwin, & Summers (1977)'s relative/normed chi-square ($\chi 2/df$) which adjusts for sample size can be adopted with recommended quotient values lying from as high as 5.0 (Wheaton et al, 1977) to as low as 2.0 (Tabachnick & Fidell, 2007).
2. **The Root Mean Square-Error of Approximation (RMSEA)**: Values at or less than .05 indicate close model fit, which is customarily considered acceptable with corresponding 90% confidence intervals (Worthington & Whittaker, 2006).
3. **Goodness-of-fit statistic (GFI)**, the adjusted goodness-of-fit statistic (AGFI): The values for the GFI and AGFI range between 0 and 1, and it is generally accepted that values of 0.90 or greater indicate well-fitting models. These two fit indices are used with other indices to establish model fit. (Hooper et al., 2012).
4. **Root mean square residual (RMR) and Standardized root mean square residual (SRMR):** Values range from zero to 1.0 with well-fitting models obtaining values less than .05 (Byrne, 1998), however values as high as 0.08 are deemed acceptable (Hu & Bentler, 1999).

Incremental Fit Indices

Incremental fit indices measure the improvement in a model's fit to the data by comparing a specific structural equation model to a baseline structural equation model that is said to fit the data poorly. They determine the relative position of model fit on a continuum that ranges from worst fit (i.e., no relationships in the data) to perfect fit. Examples:

Normed-fit index or Tucker–Lewis index (NFI or TLI): Values range between 0 and 1 values $\geq .95$ indicate a good fit (Hu & Bentler, 1999).

Comparative fit index (CFI) a value of CFI ≥ 0.95 is presently recognized as indicative of good fit (Hu & Bentler, 1999).

Parsimony Fit Indices

Parsimonious fit measures determine whether the impact of adding additional parameters on model fit is worth the decrease in degrees of freedom. Examples include:

Parsimony Goodness-of-Fit Index (PGFI) and the Parsimonious Normed Fit Index (PNFI) are commonly used. Both indices range from 0 to 1. While no threshold levels have been recommended for these indices (Crockett, 2012).

Minimum Number of Model Fit to be reported to Support Model Fit

Kline (2005) has suggested a minimum collection of fit indices to report which consist of (a) the chi-

square test statistic with corresponding degrees of freedom and level of significance, (b) the RMSEA with its corresponding 90% confidence interval, (c) the Comparative Fit Index and (d) the SRMR. (Bentler, 1995).

Hu and Bentler (1999) recommended a two-index combination strategy when reporting findings in SEM. (a) NNFI (TLI) and SRMR: NNFI of 0.96 or higher and an SRMR of .09 or lower. (b) RMSEA and SRMR: RMSEA of 0.06 or lower and a SRMR of 0.09 or lower. (c) NNFI (TLI) and SRMR: NNFI of 0.96 or higher and an SRMR of .09 or lower. (d) CFI and SRMR: CFI of .96 or higher and a SRMR of 0.09 or lower.

Detailed Explanation of the Most Commonly Used Model Indices

1. The Chi-Square Test Statistic

The chi-square test indicates the amount of difference between expected and observed covariance matrices. A chi-square value close to zero and a chi-square p-value greater than 0.05 indicate that there is little difference between the expected and observed covariance matrices, which is one indicator of good fit. A non-significant X^2 value, therefore, indicates that the theoretical model covariance matrix and the sample covariance matrix are similar.

The X^2 goodness-of-fit test is, however, sensitive to violations of the assumptions of multivariate normality and sample size. In large samples (i.e., over 200), the model chi-square statistic is nearly always statistically significant (Kline, 2005) which may lead to the rejection of a model with a good fit. Using larger sample sizes risks making a Type I error and concluding that a significant difference exists between the theoretical model covariance matrix and the sample covariance matrix, when in fact the two matrices are similar (Kelloway, 1998) While in small samples under 100 a nonsignificant probability level is nearly always indicated which may lead to accepting even a model with a poor fit. A multivariate nonnormality in the data can inflate X^2 statistics since the test assumes multivariate normality which may result in model rejections even when the model is properly specified (Crockett, 2012).

To address the limitations inherent in a chi-square test, it is often preferred to evaluate model fit based on other fit statistics (Jo¨reskog & So¨rbom, 2006; Kelloway, 1998). Secondary a relative/normed chi-square ($\chi 2/df$) statistic which minimizes the impact of sample size on the Model Chi-Square may offer an acceptable alternative. Thirdly, data should be screened for nonnormality and outliers before analysis.

2. The Root Mean Square-Error of Approximation (RMSEA)

The RMSEA evaluates the extent to which the proposed model deviates from the observed data. It chooses the model with the lesser number of parameters based on the analysis of residuals by using the square root of the mean-squared differences between the elements contained in the theoretical model covariance matrix and the sample covariance matrix (Crockett, 2012).

3. The Comparative Fit Index (CFI)

The Comparative Fit Index assesses the overall improvement of a proposed model over an independence model where the observed variables are uncorrelated (Byrne, 2006). It is least affected by sample size enabling it to perform well even when the sample size is small (Tabachnick & Fidell, 2007).

4. Non-Normed Fit Index (NNFI)/ The Tucker-Lewis Index (TLI)

The NNFI, also known as TLI determines the percentage of improvement in the theoretical model's fit as compared to the baseline model by adjusting for the degrees of freedom in the model. The value of the NNFI can indicate poor fit despite other statistics pointing towards good fit in situations where small samples are used. (Bentler, 1990; Kline, 2005; Tabachnick & Fidell, 2007).

5. Root Mean Square Residual (RMR) and Standardized Root Mean Square Residual (SRMR)

The RMR and the SRMR are the square roots of the difference between the residuals of the sample covariance matrix and the hypothesized covariance model. The range of the RMR is calculated based on the scales of each indicator. The RMR works well with scales that have homogenous levels of items e.g., all items scored on 1-3 without the same scale having other sections to be scored with another level e.g., l-5. The SRMR would naturally neutralize this problem and is therefore much more meaningful to interpret (Hooper et al., 2012).

Recommendations

University curricular developers should consider integrating more research courses into both master's and doctoral level curricula in which postgraduate students may obtain a general understanding of factor analysis (exploratory, confirmatory) and Structural Equation Modeling (SEM) statistical techniques, as well as learn how to interpret the results of such studies and apply them to practice. We can take a similar approach to the Universities where graduate programs in measurement, statistics, and psychometrics are taught (Anunciação, 2018). As a result, postgraduate students may be more informed consumers of SEM research, could better evaluate the quality of existing research, and develop rigorous research agendas (Crockett, 2012).

Conclusion

Academic research projects are commonly carried out in different disciplines all over the world, including in Uganda. While a lot of literature can be found on scale theory and development, instruments with incomplete scale development are often being used by researchers in data gathering when measuring the attributes of their study, and as such tarnishing the credibility of their research findings. To save the credibility of their research findings literature suggest that researchers either utilize already developed and validated instruments or commit to the scientifically rigorous and comprehensive process of developing an instrument for their research project.

We hope that this article will motivate university curricular developers to address their own knowledge and practical gaps in scale development and evaluation based on the fundamental concepts in psychometric theory, and accept, include, and apply, in graduate training in Uganda's higher education, all the recommendations in this article of observational and documentary study

References

Abbo, C., Ekblad, S., Waako, P., Okello, E., & Musisi, S. (2009). The prevalence and severity of mental illnesses handled by traditional healers in two districts in Uganda. *African Health Sciences*, Vol. 9, Special Issue 1, S16-S22.

Ahia, C. E., & Bradley, R. W. (1984). Assessment of secondary school student needs in Kwara State, Nigeria. *International Journal for Advancement of Counseling, 7,* 149-157.

Ajzen I. (1985) From intentions to actions: a theory of planned behavior. In: Action Control SSSP Springer Series in Social Psychology. Springer. p. 11–39.

American Psychological Association (APA). (1995). *Standards for educational and psychological testing*. Author.

Anunciação, L. (2018). An Overview of the History and Methodological Aspects of Psychometrics-History and Methodological aspects of Psychometrics. Journal for ReAttach Therapy and Developmental Diversities. Aug 15; 1(1):44-58.

Bai Y, Peng C-YJ, & Fly, A.D. (2008). Validation of a short questionnaire to assess mothers' perception of workplace breastfeeding support. *J Acad Nutr Diet*, 108:1221–5.

Hirani, S.A.A, Karmaliani R, Christie T., & Rafique, G. (2013). Perceived Breastfeeding Support Assessment Tool (PBSAT): development and testing of psychometric properties with Pakistani urban working mothers. *Midwifery*, 29:599–607.

Barrett, G. V. (1972). New research models of the future for industrial and organizational psychology. *Personnel Psychology, 25*, 1-17.

Barrett, P. (2007), "Structural Equation Modelling: Adjudging Model Fit," *Personality and Individual Differences, 42 (5),* 815-24.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, *3*, 77-85.

Bartlett, J. E., Kortrilik, J. W., & Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*(1), 44-50.

Bentler, P.M., & Chou, C. (1987) Practical Issues in Structural Modeling. *Sociological Methods and Research, 16*, 78- 117.

Bentler, P. M. (1995). *EQS: Structural equations program manual*. Multivariate Software

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bjorner, J. B., Kosinski, M., & Ware, J. E. (2005). Computerized adaptive testing and item banking. In Fayers, P.M., Hays, R.D. (Eds.). *Assessing Quality of Life*. Oxford University Press.

Boateng, G. O., Neilands, T.B., Frongillo, E. A., Melgar-Quiñonez, H.R., &. Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front Public Health, 11*(6):149. doi: 10.3389/fpubh.2018.00149

Boynton P.M., & Greenhalgh, T. (2004). Selecting, designing, and developing your questionnaire. *BMJ*, ;*328:13*,12–5.

Brown, R. (2011). *Prejudice: Its social psychology*. John Wiley & Sons.

Buchanan, R. D., & Finch, S. J. (2005). History of psychometrics. *Encyclopedia of statistics in behavioral science*. John Wiley & Sons, Ltd.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Lawrence Erlbaum Associates Publishers.

Byrne, B. M. (1989). *A primer of LISREL*. Springer-Verlag.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Erlbaum.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Child, D. (1990). *The essentials of factor analysis* (2nd ed*)*. Cassel Educational Limited.

Cooley, W. W. (1978). Explanatory Observational Studies. *Educational Researcher*, *7(9),* 9-15.

Crockett, S. A. (2012). A five-step guide to conducting SEM analysis in counseling research. *Counseling Outcome Research and Evaluation, 3*(1), 30-47.

DeVellis, R.F. (2012). *Scale Development: Theory and Application*. Sage Publications.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power

3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* Springer-Verlag.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction.* Sage Publications, Inc.

Hatcher, L. (1994) A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. SAS Institute, Inc.

Hinkin, T. R. (1995). A review of scale development in the study of behavior in organizations. *Journal of Management, 21*,967-988.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 2*(1), 104-121.

Hirani S. A. A., Karmaliani, R, Christie, T., & Rafique, G. (2013). Perceived Breastfeeding Support Assessment Tool (PBSAT): development and testing of psychometric properties with Pakistani urban working mothers. *Midwifery*, 29:599–607.

Holtzman, S. & Vezzu, S. (2011). Confirmatory factor analysis and structural equation modeling of noncognitive assessments using PROC CALIS. Northeast SAS Users Group (NESUG), 2011 proceedings, 11-14.

Hooper, D., Coughlan, J., & Mullen, M. R. (2012). *Structural equation modelling: Guidelines for determining model fit*. Retrieved March 30, 2013, www.ejbrm.com

Hu, L.T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jones, L. V., & Thissen, D. (2007). A History of Psychometrics. *Handbook of Statistics, 26*, 1-20.

Joreskog, K. G., & Sorbom, D. A. (2006). LISREL 8.54 and PRELIS 2.54. Scientific Software.

Jöreskog, K.G., & Sörbom, D. (2004). LISREL 8.7. Scientific Software International, Inc.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187–200. https://doi.org/10.1007/BF02289233

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Sage.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford Press.

Marsh, H.W., Hau, K.T., & Wen, Z. (2004), "In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's Findings ". *Structural Equation Modeling, 11 (3),* 320-41.

McDonald, R.P. (1978), "A simple comprehensive model for the analysis of covariance structures," *British Journal of Mathematical and Statistical Psychology,* 37, 234-251.

Nakigudde, J., Musisi, S., Ehnvall, A., Airaksinen, E., & Agren, H. (2009). Adaptation of the multidimensional scale of perceived social support in a Ugandan setting. *African Health Services, 9*, S35-S41.

Nakimuli-Mpungu, E., Musisi, S., Katabira, E., Nachega. J., & Brass, J. (2011). Prevalence and factors associated with depressive disorders in an HIV+ rural patient population in southern Uganda. *Journal of Affective Disorders, 135*(1) 160-167.

Nsereko, D. N., Musisi, S., & Holtzman, S. (2014). Evaluation of psychosocial problems among African university students in Uganda: Development and validation of a screening instrument. *Psychology Research*, *2*(4), 112-131.

Osborne, J. W. (2010). Challenges for quantitative psychology and measurement in the 21st century. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2010.00001.

Ovuga, E. (2005). *Depression and suicidal behavior in Uganda: Validating the response inventory for*

stressful life events (RISLE). Unpublished Doctoral thesis, Karolinska Institute, Sweden and Makerere University, Uganda.

Ovuga, E., Boardman, J., & Wasserman, D. (2006). Undergraduate student mental health at Makerere University, Uganda. *World Psychiatry, 5*(1), 51-52.

Price, J. L., & Mueller, C. W. (1986). *Handbook of organizational measurement*. Pitman.

Reijneveld, S. A., Vogels, A. G. C., Brugman, E., Van Ede, J., Verhulst, F. C., & Verloove-Vanhorick, S. P. (2003). Early detection of psychosocial problems in adolescent. How useful is the Dutch short indicative questionnaire (KIVPA)? *European Journal of Public Health, 13*, 152-159.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287-297.

Savalei, V., & Rhemtulla, M. (2017). Normal Theory GLS Estimator for Missing Data: An Application to Item-Level Missing Data and a Comparison to Two-Stage ML. *Front. Psychol.* 8:767. doi: 10.3389/fpsyg.2017.00767

Schoenfeldt, L. F. (1984). Psychometric properties of organizational research instruments. In T. S. Bateman & G. R. Ferris (Eds.), *Method & analysis in organizational research* (pp. 68- 80). Reston.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research; 99(6),* 323-337.

Senyonyi, R. M., Ochieng, L. A., & Sells, J. (2012). The development of professional counseling in Uganda: Current status and future trends. *Journal of Counseling & Development, 90*, 500-505.

Stone, E. (1978). *Research methods in organizational behavior.* Scott, Foresman.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Harper & Row.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

Thompson, D. W., Loesch, L. C., & Seraphine, A. E. (2003). Development of an Instrument to Assess the Counseling Needs of Elementary School Students. *Professional School Counseling*, *7*(1), 35-39.

van der Linden, W. J., & Hambleton, R. K. (1997). (Eds.). *Handbook of modern item response theory*. Springer-Verlag.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8*(1*), 84-136.*

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2021). Values in Psychometrics. *Perspectives on Psychological Science.*

https://doi.org/10.1177/17456916211014183

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806-838.