Janabhawana Research Journal
Vol. 3, Issue No. 1, July 2024, Page: 35-49
ISSN: 2505-0621

A Pilot Study Approach to Assessing the Reliability and Validity of Relevancy and Efficacy Survey Scale

Deb Bahadur Chhetri [1] & Bishnu Khanal [2]*

Abstract

This paper is based on the pilot study that aims to assess the reliability and validity of the attitude measurement scale towards the relevance of mathematical knowledge of real analysis for secondary-level mathematics instruction. Research tool validation and reliability estimation are challenging due to the complexity of its procedure. This study establishes procedural guidance for piloting and tool validation. The survey method was used to conduct the pilot study. Thirty participants filled in the 45 indicators of relevance measurement and 10 mathematical efficacy measurement indicators. The reliability was assessed with Cronbach's Alpha formula while the validity of the questionnaire was examined through item-total correlation with each item by using Pearson's formula. The content validity index was also calculated based on the expert judgment method. The reliability indicator was found to be more than 0.70 and the validity indicator falls in the acceptable range. Therefore, developed scale is valid and reliable to measure the teachers' attitudes on the relevancy of mathematical knowledge of real analysis of school's mathematics instruction based on pilot study results. Furthermore, this study guides the pilot study procedure for survey research and especially benefits those looking to validate their Likert-type survey scale in any research.

Keywords: reliability, survey scale, validation process

Article information

Received: 24-05-2024 Reviewed: 07-06-2024 Revised: 22-06-2024 Accepted: 27-06-2024

** Corresponding Author's Email: dev.chhetri@dmc.tu.edu.np*
Orcid: https://orcid.org/0000-0002-1860-1395

Cite this article as:

Chhetri, D., & Khanal, B. (2024). A pilot study approach to assessing the reliability and validity of relevancy and efficacy survey scale. *Janabhawana Research Journal, 3*(1), 35-49. https://doi.org/10.3126/jrj.v3i1.68384

[1] *Teacher at Dhawalagiri Multiple Campus, TU, Nepal*
[2] *Associate professor of Mathematics Education, Faculty of Education, TU, Nepal, https://orcid.org/0000-0002-3304-7695*

Introduction

The courses of the teacher preparation program become relevant when teachers utilize these courses in their professional activities and help the teacher to increase their professional activities. Mathematics teacher preparation courses in almost all Universities of the world have introduced advanced mathematics (Schmidt et al., 2013). Different universities of Nepal like TU have introduced advanced mathematics courses like real analysis in their different programmes (Panthi & Jha, 2016). It is expected that these courses are relevant for being a secondary-level mathematics teacher. Besides such expectation, there is not sufficient evidence about its relevancy. Therefore, a survey instrument is prepared to measure its relevancy. The instruments are prepared based on prior scholarly work and theories like mathematical knowledge for teaching (Ball et al., 2008) and the training evaluation model developed by Kirkpatrick & Kirkpatrick (2014). These are major theoretical bases for instrument preparation which focuses on the connectivity of content and its utilization in practice and that can be measured from the response of teachers.

A reliable and valid instrument is required to measure the desired construct which indicates the relevancy from the perspective of the teachers.

Reliability and validity are essential cornerstones in the development and validation of research instruments. Reliability indicates consistency and stability and validity refers to the accuracy and appropriateness of measuring instruments. The ability of a survey instrument to deliver consistent and repeatable data is referred to as reliability (Mallinger & Hanson, 2020). The result from the survey scale should be stable across time, items, and groups and the variation in numerical measures is presumed the result of the error (Nunnally, 1978). The common method of calculating reliability indicators on a survey scale is Cronbach Alpha (Ghazali, 2016). This Cronbach Alpha can be calculated by using the software. Validity is measured based on its type and requirements. Especially Pearson's' correlation between each item and the item total is calculated (Suhartini et al., 2021) during a pilot study for construct validity and using expert judgment for content validation (Aithal & Aithal, 2020). These methods are useful to check the reliability and validity of survey instruments.

Survey tool development and estimation of the reliability and validity of the research tools are important in academic research. It helps to ensure the developed tools consistently measure construct and measure those constructs which are supposed to be measured before truly launching survey tools we have to verify it in a small group of target population and refine if needed. This study aims at developing and piloting survey tools to assess the relevance of real analysis knowledge for school mathematics instruction, ensuring their reliability and validity while exploring all processes practically.

Literature Review

Theoretical Foundation of Survey Tool Development

Survey research tool development is a more challenging task. Regarding the development of survey tools, several steps are required. Each step contributes to the overall quality and efficacy of the instrument.

Several theoretical frameworks underpin the steps required to develop research survey tools. Construct Validity Theory, proposed by Cronbach and Meehl (1955), underscores the need to define constructs clearly to ensure that survey items measure what they are intended to. Item Response Theory (IRT), developed by Lord and Novick (1968), supports the design of survey items by focusing on their relationship with latent traits and ensuring they reflect the underlying construct accurately. Formative Evaluation Theory, articulated by Scriven (1991), highlights the importance of pilot testing as a means of refining and improving survey instruments through iterative feedback. Classical Test Theory (CTT), rooted in Spearman's work (1904), provides methods for assessing reliability and validity, with Cronbach's Alpha being a key measure of internal consistency. Validity Theory, according to Messick (1989), emphasizes the comprehensive analysis of content, criterion-related, and construct validity to ensure accurate measurement. Finally, Iterative Design Theory, described by Schon (1983), advocates for an iterative process involving multiple revisions based on feedback and evaluation, enhancing the survey tool's effectiveness through continuous refinement.

Reliability and Validity of Scale

To ensure the accuracy, consistency, and efficacy of the tools, reliability and validity play a critical role in their development and application. The method of estimating reliability and validity and their value and their meaning for the adequacy of the tool are discussed in the upcoming section.

The reliability of an instrument refers to its stability and consistency over time and similar samples. The reliability indicator is considered as the consistency of the instrument. While a reliability coefficient may be notably high and satisfactory, it does not ensure precise measurement of the construct (Hair et al., 2010). Thus reliability is required in survey instruments. Internal consistency, test-retest reliability, parallel form reliability and inter-rater reliability are major methods of examining the reliability of the questionnaire.

In the survey instrument, the commonly used method is measuring the internal consistency by calculating Cronbach's Alpha coefficient (Aithal & Aithal, 2020). The standard value of Cronbach's Alpha coefficient lies between 0.70-0.90 for adequate internal consistency (Robinson, 2009; Hamed, 2016). Additionally, a widely acknowledged guideline suggests that Cronbach's alpha (α) within the range of 0.6 to 0.7 signifies an acceptable level of reliability, while a value of 0.8 or higher indicates a very good level. Nevertheless, caution

is advised when alpha values surpass 0.95, as such elevated values might suggest redundancy in the measurements (Hulin et al., 2001). Thus, Cronbach Alpha is calculated to check the internal consistency of the survey instrument and its acceptable value falls under 0.70-0.95.

The validation process of a survey questionnaire involves examining the survey questions to assess their reliability. Given the numerous intricate factors that can impact the reliability of questions in a questionnaire, survey validation becomes a complex procedure. The determination of a questionnaire's validity relies on understanding its intended measurement purpose. There are various methods of testing the validity of survey instruments. Face validity, content validity and construct validity are measured during the pilot study.

Face validity of a questionnaire refers to the degree to which test items or questionnaire questions appear to be assessing the idea they are intended to measure and it can be established by a judging expert panel (Desai & Pater, 2020). Thus it is not statistical techniques to improve by taking the help of expert judgment.

Content validity refers to the quality of measurement instruments that ensure items adequately cover the construct's domain. An expert evaluation can be performed to check the content validity. The content validation indexes need to be calculated by using statistical methods. The acceptable range of the content validity index depends on the number of experts involved in the judgment process. If only two experts are involved in judgment then the acceptable content validity index (CVI) should be at least 0.80 (Davis, 1992). If three to five experts are engaged in the assessment, the acceptable content validity index (CVI) should be 1. When involving at least six experts, the CVI should be at least 0.83 (Polit & Beck, 2006; Polit et. al. 2007). Moreover, for six to eight experts, the CVI should be a minimum of 0.83, and for at least nine experts, it should be at least 0.78 (Lynn, 1986). Hence content validity index depends upon the number of experts involved in judgment in scale.

Construct validity refers to the quality of the questionnaire where it measures those constructs that are intended to be measured. The process of determining if a measurement instrument, such as a questionnaire, is measuring the theoretical construct or notion it is supposed to measure is known as construct validity. Pearson's' correlation between each item and the item total is calculated (Machuca et. al., 2015; Suhartini et al., 2021) to find the construct validity index. When comparing this with Pearson's correlation between the item and the total score, a higher Pearson's correlation would indicate that the item is contributing significantly to the overall scale score.

Apart from the earlier discussed methods for evaluating the reliability and validity of survey instruments, the Fornell-Larcker criterion (Franke & Sarstedt, 2019) and the Heterotrait-Monotrait (HTMT) ratio (Henseler & Sarstedt, 2015) provide significant approaches in the field of construct validation. The first method compares the square root of

the average variance extracted (AVE) with the correlation between the construct where each correlation should be less than the square root of AVE for validity. The HTMT ratio focuses on divergent validity. By comparing the correlations within a construct (monotrait) with those across different constructs (heterotrait), researchers gain insights into whether the measurements are distinct. A ratio lower than a predefined threshold (commonly 0.85) suggests satisfactory discriminant validity. However, these are utilised for structural equation modelling and use comparatively large sample sizes in the comparison of piloting sample sizes.

Hence, Cronbach's Alpha and the item-total correlation method for reliability and validity were used to check the reliability and validity of the developed survey instrument.

Sample size for Pilot Study

A pilot study serves as a preliminary and small-scale investigation conducted before the main research project. As its primary goal is not hypothesis testing, there is often no requirement to calculate the sample size for the pilot study. Different small sample sizes can be found for a pilot study. A pilot study by Omar et al., (2017) utilizes a total sample size of 24 where a total range of sample may be 10-40 for the pilot study (Lewis et al., 2021). Additionally, Julious (2005) suggests 12 for a pilot study. Hence 30 respondents are considered for the pilot study.

The practical application and procedural approaches are often not covered in a single article, but they can be consolidated into a single questionnaire. Therefore, this study focuses on both theoretical and practical approaches to develop and pilot tools, ensuring their reliability and validity.

Method

The Likert-type survey tools on the relevance of mathematical knowledge of real analysis were developed based on the literature and theoretical guidelines. Regarding tool development Taherdoost (2022) outlines a structured approach to designing an effective questionnaire for research. The process begins with defining the research objectives and identifying the target population. Next, researchers develop questions that align with the objectives and design the questionnaire format. A pre-test with a small group helps refine the tool before its full administration. After distributing the finalized questionnaire, researchers collect and analyze the data to draw conclusions and report findings. This systematic approach ensures the questionnaire is clear, relevant, and reliable. Especially this guideline of tool development was utilized. The steps adopted to develop tools are discussed below.

1 Developing Framework

The framework for the questionnaire was prepared based on the theories. Especially, relevancy framework was especially adapted from the training evaluation model developed

by Kirkpatrick & Kirkpatrick (2014). Additionally, the mathematics teacher development framework developed by Ball et al., (2008) was used. First theory guides the relevancy measure and second theory to add the what type of knowledge is required for being mathematics teacher. Based on this frame the content connection, pedagogical application and mathematical proficiency development were considered for the relevancy measurement scheme. Additionally, the teacher efficacy measurement scale was prepared based on guidelines by Tschanen and Hoy (2001).

2 Item development

The questionnaire items are developed based on the literature and the framework. fifteen items of each domain of the relevancy measurement scale were prepared. Additionally, fifteen items of the efficacy measurement scale were prepared in the initial stage. The survey tool designed to assess secondary-level mathematics teachers' perspectives on the relevance of real analysis mathematical knowledge emphasizes aspects such as content integration, pedagogical application, and skill enhancement. Furthermore, the measurement of mathematics teachers' efficacy in the instrument was crafted by drawing upon existing literature and relevant theoretical frameworks.

3 Item screening, improving and preparing for piloting

First, a colleague discussion was conducted in September 2023 for initial improvement. Language correction and item correction were made at the suggestion of a colleague.

Table 1

Nature of Scale

Category	Domains	Number of items	Response Nature
Relevancy measuring scale	Content Connection Pedagogical Application Enhancing mathematical proficiency	15 (1-15) 13(16-28) 17(29-45)	SA= Strongly agree, A= Agree, N= Neural, D= Disagree, SD= strongly disagree
Efficacy measurement Scale	Creation Fluency Justification	3 (1,2,7) 3(4,5,8) 4 (3,6,9,10)	Rate1 to 5 where: minimum 1 and maximum 5) based on ability

The prepared items were sent to the university professor based on their permission and acceptance after telephone communication. After their suggestion, it was improved. The initial draft of the questionnaire was created and subsequently submitted to experts for a face validity check. Corrections were then implemented based on the expert's suggestions, and this iterative process was repeated five times to refine and enhance the questionnaire.

The relevancy measurement questionnaire was improved and prepared based on five-point Likert scales whereas the mathematical efficacy measurement scale was based on a self-rating scheme from one –to five. The structure of the questionnaire is shown in Table 1.

4 Pretesting (piloting)

The improved scale was introduced and tested in a pilot study conducted in October 2023. Five experts were contacted for content validity judgment. Only four accepted my request for their voluntary work. Consent was taken from the experts. The questionnaire was distributed to the experts online via Google Forms, rating their responses based on a rating scale: 1 = item is not relevant to measure the domain; 2 = item is somewhat relevant to measure the domain; 3 = item is quite relevant to measure the domain and 4 = item is highly relevant to measure the domain. Two experts textured the assessed report on content validity. After collecting their responses, the Content Validity Index (CVI) was calculated to gauge the extent of content validity.

Following the content validity check, the questionnaire was administered to secondary-level mathematics teachers, and their responses were gathered through both online and physical modes. A total of 30 participants were selected for the pilot study. After collecting the responses, the data was entered into the SPSS software for an analysis of reliability and validity. Internal consistency was examined by calculating Cronbach's Alpha, providing insights into the reliability of the questionnaire. Additionally, the construct validity was assessed through the calculation of item-total Pearson's correlation using the SPSS software. These statistical analyses were conducted to ensure the strength and validity of the instrument in measuring the intended constructs.

Result and Discussions

The result of the process and indicators of checking reliability and validity are expressed in the following title separately.

Content Validity Index (CVI)

Before calculating the content validity index the relevance rating was recorded as (1= item is not relevant to the measured domain, 2= item is somewhat relevant to measure the domain, 3= the item is quite relevant to the measured domain, 4= item is highly relevant to measured domain) as given in the instruction for the response as shown in Table 2

a) Calculation of agreement: Agreement is determined by counting the number of experts who assigned a rating of 3 or 4 to an item. Table 2 illustrates this process. For instance, in the case of item 1, the agreement score is 2, as two experts gave a rating of 4. Conversely, for item 8, the agreement score is 1, as only one expert rated it as 4, while expert A provided a rating of 1.

Table 2

Content Validity Index of Relevancy Measurement Scale

	Expert rating on relevance				CVI Indicators	
Item	Expert A	Expert B	Expert agreement	I-CVI	UA	I-CVI
1	4	4	2	1	1	1
2	4	4	2	1	1	1
3	4	4	2	1	1	1
4	4	4	2	1	1	1
5	4	4	2	1	1	1
6	4	4	2	1	1	1
7	4	4	2	1	0	1
8	2	4	1	0.5	1	0.5
9	4	3	2	1	1	1
10	4	3	2	1	1	1
11	4	4	2	1	1	1
12	4	4	2	1	1	1
13	4	4	2	1	1	1
14	4	4	2	1	1	1
15	4	4	2	1	1	1
	P=0.93	P= 1	SCVI Average =0.965	SCVI average = 0.97	S-CVI/UA = 0.93	S-CVI/UA = 0.96

The expert agreement refers to the number of experts who agreed relevant to the item. There are the experts who rated 3 or 4 in the item.

b) I-CVI: It is calculated by using the formula $ICVI = \frac{Value\ of\ expert\ agreeemnt}{Total\ number\ of\ expert}$. By using this formula I-CVI of item 1 = $\frac{2}{2} = 1$ but the I-CVI of item 8 is:$\frac{1}{2} = 0.5$. Similarly, all I-CVI are calculated as a ratio of the value of the expert agreement and the number of experts.

c) Universal Agreement (UA): It is obtained by placing the value 1 if all experts rated the item 3 or 4 which indicates the item is relevant universally otherwise score 0 which indicates the item is not fully relevant from an expert perspective. For example, UA of item 1 is 1 where both experts rated 4, similarly in item 9 the rated value is 3 and 4. But in item 8, UA is 0 because one expert rated 2.

d) SCVI average: It is calculated based on the number of items and total number of experts. The calculation based on the number of items is: S-CVIave = $\frac{I-CVI\ score}{Total\ number\ of\ item}$. For example **SCVI average** = (1+1+1+1+1+1+1+.5+1+1+1+1+1+1+1)/15=0.97. Additionally, **SCVI** average based on expert is: SCVI $_{average}$ = $\frac{sum\ of\ relevance\ proportion}{Total\ number\ \ expert}$. For example

SCVI average = (0.93+1)/2 =0.965.

E) S-CVI/UA: It is the average of UA score of all items. It is calculated by dividing the sum of UA by the total number of items. Eg. (1+1+1+1+1+1+0+1+1+1+1+1+1+1+1)/15 =0.93

Table 3

Content Validity Index of Relevancy Measurement Scale

Part	Domains	Indicators		
Relevance		SCVI-Average based on expert	SCVI –Average based on item	S-CVI/UA
	Content Connection	0.96	0.97	0.93
	Pedagogical Application	1	1	1
	Proficiency development	0.97	0.97	1
Efficacy for teaching	Creating and Evaluation	1	1	1
	Fluency	1	1	1
	Logical fluency	1	1	1

The calculated value of the content validity index of each domain questionnaire of relevancy measurement scale and mathematics teaching efficacy scale exceeded the threshold for two expert judgments of 0.80 recommended by Davis (1998). Consequently, the content validity of both the relevance measurement questionnaire and the teaching efficacy measurement questionnaire is strong.

Reliability

Several techniques, including parallel form test-retest, Kuder-Richardson, and Cronbach's Alpha, are commonly used to assess questionnaire reliability. In this study, the Cronbach's Alpha method was specifically employed as suggested by Aithal and Aithal (2020) and the corresponding result is detailed in Table 4. The Cronbach's Alpha values for both the total scale and individual items of both scales fall within the recommended threshold range of 0.70-0.95, as outlined by Robinson (2009) and Hamed (2016) indicating sufficient internal consistency for both the scale and its items.

Hence, the developed scales have a strong level of reliability index because the pilot study result satisfied the criteria for internal consistency.

The analysis covered inter-item relation, item-total correlation, Cronbach alpha values (when the respective item is deleted), and item-total statistics for item analysis. The inter-item statistics for each item ranged from 0.307 to 0.674 and were all positive values, an indication that items fit together conceptually (DeVon et al., 2007). Low inter-item statistics would suggest non-discriminating items, while high inter-item statistics would show that each item is not adding something distinct to the concept, implying multidimensionality. Nonetheless, there was not a single instance in the data set where the inter-item correlation and Cronbach's alpha coefficient values were both outside of an acceptable range. Table 4

demonstrated that the item-total correlation amongst items was found to be acceptable, as all items had a good correlation with other items.

Table 4

Reliability of Relevance and Teaching Efficacy Scale

Scale	Domains	Crohbach's Alpha	Range of adequacy	Interpretation
Relevance of Mathematical Knowledge	Content Connection	0.826	0.70-0.95	Reliability is adequate
	Pedagogical Application	0.845		
	Enhancing Proficiency	0.892		
	Total item	0.943		
Mathematics teaching efficacy	Creation	0.740	0.70-0.95	Reliability is adequate
	Procedural fluency	0.792		
	Reasoning	0.710		
	Total item	0.897		

Total item Statistics of Relevance Measurement Scale

Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	0.619	0.941	16	0.373	0.943	31	0.510	0.942
2	0.480	0.942	17	0.326	0.943	32	0.559	0.942
3	0.307	0.943	18	0.653	0.941	33	0.500	0.942
4	0.674	0.941	19	0.427	0.942	34	0.522	0.942
5	0.354	0.943	20	0.597	0.941	35	0.632	0.942
6	0.565	0.942	21	0.564	0.942	36	0.512	0.942
7	0.593	0.941	22	0.502	0.942	37	0.538	0.942
8	0.379	0.943	23	0.645	0.941	38	0.504	0.942
9	0.640	0.941	24	0.504	0.942	39	0.414	0.943
10	0.354	0.943	25	0.488	0.942	40	0.605	0.942
11	0.407	0.943	26	0.670	0.941	41	0.556	0.942
12	0.440	0.942	27	0.399	0.943	42	0.623	0.941
13	0.608	0.941	28	0.424	0.943	43	0.633	0.941
14	0.388	0.943	29	0.500	0.942	44	0.568	0.942
15	0.402	0.943	30	0.508	0.942	45	0.467	0.942

Total item Statistics Teaching Efficacy Measurement Scale

Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	0.764	0.878
2	0.733	0.881
3	0.521	0.894
4	0.658	0.886
5	0.679	0.885
6	0.584	0.891
7	0.543	0.893
8	0.695	0.884
9	0.601	0.890
10	0.676	0.885

The item-total correlations were seen to be within .30 to .70 and can be considered acceptable (deVaus, 2004) for all the items with total scores in the range of .30 and .70 (Carmines & Zeller 1974). Based on this result it is concluded that the reliability of the total scale, domains and each item are good there reliable for further use.

Construct Validity

Pearson's correlation between the sum of items and each item was calculated to assess construct validity, following the recommendations of Machuca et al., (2015) and Suhartini et al., (2021) by using SPSS software. The result was found as shown in Table 5.

Table 5

Item total Correlation of the Scale

Validity coefficient of Relevance Measurement Scale				Teaching Efficacy Items' Validity	
Item	Correlation	items	r	Item	r
1	.645**	25	.518**	1	.822**
2	.532**	26	.657**	2	.790**
3	.369*	27	.484**	3	.599**
4	.692**	28	.464**	4	.730**
5	.374*	29	.520**	5	.760**
6	.607**	30	.538**	6	.678**
7	.617**	31	.520**	7	.632**
8	.381*	32	.575**	8	.775**
9	.664**	33	.541**	9	.671**
10	0.357	34	.560**	10	.748**
11	.424*	35	.640**		
12	.421*	36	.568**		
13	.610**	37	.574**		
14	.428*	38	.537**		
15	.427*	39	.439*		
16	.415*	40	.636**		
17	.370*	41	.560**		
18	.667**	42	.653**		
19	.468**	43	.654**		
20	.636**	44	.583**		
21	.567**	45	.484**		
22	.544**				
23	.652**				
24	.571**				

**Correlation significant at 0.01 level and *correlation significant at 0.05 level

The validity of each item was determined from the significance of its correlation with the sum of the score (item-total correlation) and compared with the significance value of Pearson's correlation (0.361).

Except for item 10, Pearson's correlation coefficient was positive and significant for all items. Specifically, the correlation coefficients for items 3, 5, 8, 10, and 17 ranged between 0.35 and 0.39, which, according to Schober et al., (2018), is considered low but still significant. Meanwhile, correlations for the remaining items were classified as moderate (0.40-0.69) and strong (0.70-0.89) based on the criteria by Schober et al., (2018). Overall, these findings suggest that the items exhibit adequate validity for further utilization.

Conclusion

Among the different methods of calculating the reliability and validity of the questionnaire, Cronbach's Alpha is suitable for checking reliability and content validity index (CVI) and item-total correlation suitable during the pilot study with a sample of range 10-40. The result of this pilot study meets the criteria of reliability and validity.

Hence the 45-item relevance measurement scale with three domains: content connection, pedagogical application, and enhancing proficiency development is reliable and valid for measuring teachers' attitudes in terms of content connection, pedagogical application, and enhancing the mathematical proficiency by mathematical knowledge of real analysis. Additionally, 10 items for the teaching efficacy measurement scale with three domains: creation, procedural fluency, and reasoning are reliable and valid for measuring the mathematical efficacy of teaching.

Acknowledgments

We wish to express our sincere appreciation to the esteemed experts who graciously dedicated their time and expertise to assess the reliability and validity of the tool developed for the doctoral research conducted by Mr. Deb Bahadur Chhetri, the first author of this article. Your invaluable feedback and insightful comments have significantly contributed to enhancing the quality of this research.

We are deeply grateful to the Graduate School of Education at Tribhuvan University for their unwavering support, provision of resources, and creation of a conducive research environment in the doctoral journey.

Conflict of interest: We discovered that there is no conflict of interest

References

Aithal, A., & Aithal, P. S. (2020). Development and validation of survey questionnaire & experimental data systematical review-based statistical approach. *International Journal of Management, Technology, and Social Sciences (IJMTS), 5*(2), 233-251. https://doi.org/10.47992/IJMTS.2581.6012.0116

Ball, L. D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407. https://doi.org/10.1177/0022487108324554

Brown, J. D. (2000). What is construct validity? Shiken: JALT Testing & Evaluation. *SIG Newsletter, 4* (2), 8-12.

Carmines, E. G., & Zeller. R. A. (1974). On establishing the empirical dimensionality of theoretical terms: An analytic example. *Political Methodology, 1*(4), 75-96. https://www.jstor.org/stable/25791395

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://doi.org/10.1037/h0040957

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194-197. https://doi.org/10.1016/S0897-1897(05)80008-4

De Vaus, D. (2004). Surveys in Social Research (5th ed.). London: Routledge. https://books.google.com.np/books?hl=en&lr=&id=rnxiAgAAQBAJ&oi=fnd&pg=PP1&dq=De+Vaus,+D.+(2004).+Surveys+in+Social+Research+(5th+ed.).+London:+Routledge.&ots=6fOIGijDKT&sig=4fuUGa7diA67BLtf-j5nPey7o6U&redir_esc=y#v=onepage&q&f=false

Desai, S., & Patel, N. (2020). ABC of face validity for questionnaire. *International Journal of Pharmaceutical Sciences Review and Research, 65* (1), 164-168. http://dx.doi.org/10.47583/ijpsrr.2020.v65i01.025

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155-164.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105-123.

Franke, G. R., & Sarstedt, M. (2019). Heuristics Versus Statistics in Discriminant Validity Testing: A Comparison of Four Procedures*, Internet Research, 29*(3), 430-447. https://doi.org/10.1108/IntR-12-2017-0515

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM)*, *5*. https://doi.org/10.2139/ssrn.3205040

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A New Criterion for Assessing Discriminant Validity in Variance-based Structural Equation Modeling., *Journal of the Academy of Marketing Science, 43*(1): 115-135. https://doi.org/10.1007/s11747-014-0403-8

Hulin, C., Netemeyer, R., & Cudeck, R. (2001). Can a Reliability Coefficient Be Too High? *Journal of Consumer Psychology*, *10* (1), 55-58.

In, J. (2017). Introduction of a pilot study. *Korean Journal of Anesthesiology*, *70*(6), 601–605. https://doi.org/10.4097/kjae.2017.70.6.601

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics, 4,* 287-291https://doi.org/10.1002/pst.185

Kirkpatrick, J., & Kirkpatrick, W. (2014). The Kirkpatrick four level: a fresh look after 55 years. In *Krikpatrick Partners*. http://www.kirkpatrickpartners.com/Portals/0/Resources/White Papers/Introduction to the Kirkpatrick New World Model.pdf

Lewis, M., Bromley, K., Sutton, C. J., McCray, G., Myers, H. L., & Lancaster, G. A. (2021). Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back! *Pilot and feasibility studies*, *7*(1), 1-14. https://doi.org/10.1186/s40814-021-00770-x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382-386. https://doi.org/10.1097/00006199-198611000-00017

Machuca, C., Baker, R. S., Sufi, F., Mason, S., Barlow. A., & Robinson, P.G (2015). 10-derivation of a short form of the dentine hypersensitivity questionnaire. *Dentine Hypersensitivity,* 155-164. https://doi.org/10.1016/B978-0-12-801631-2.00010-5

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 19*, 172–190.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). *American Council on Education.*

Ghazali, H. M. (2016). A reliability and validity of an instrument to evaluate the school-based assessment system: A pilot study. *International journal of evaluation and research in education, 5*(2), 148-157. https://doi.org/10.11591/ijere.v5i2.4533

Nunnally, J.C. (1978). *Psychometric theory*. McGraw-Hill.

Omar, M., Klawonn, F., Brand, S., Stiesch, M., Krettek, C., & Eberhard, J. (2017). Transcriptome-wide high-density microarray analysis reveals differential gene transcription in periprosthetic tissue from hips with chronic periprosthetic joint infection vs aseptic loosening. *The Journal of Arthroplasty*, *32*(1), 234-240. https://doi.org/10.1016/j.arth.2016.06.036

Panthi, D., & Jha, K. (2016). A study on teaching applicable mathematics in the universities of Nepal: A study on teaching applicable mathematics in. *Kathmandu University Journal of Science, Engineering and Technology.*, *1*(5), 10–12. https://doi.org/10.3126/kuset.v4i1.2888

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497. https://doi.org/10.1002/nur.20147

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459-467. https://doi.org/10.1002/nur.20199

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology, 38*(2), 337-341. https://doi.org/10.1093/ije/dyn357

Schmidt, W., Burroughs, N., & Cogan, L. (2013). World Class Standards For Preparing Teachers of Mathematics. In *Center for the study of curriculum at Michigan State University*.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia, 126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Schon, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Scriven, M. (1991). *Evaluation Thesaurus*. Sage Publications.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293. https://doi.org/10.2307/1412107

Suhartini, R., Nurlaela, L., Wahyuningsih, U., & Prihatina, Y. I. (2021). Validity, reliability, intra-rater instrument parameter teaching factory and learning outcomes of industrial clothing. In *International Joint Conference on Arts and Humanities 2021 (IJCAH 2021)* (pp. 1230-1239). Atlantis Press. https://doi.org/10.2991/assehr.k.211223.214

Taherdoost, H. (2022). Designing a questionnaire for a research paper: A comprehensive guide to design and develop an effective questionnaire. *Asian Journal of Managerial Science, 11*(1), 8-16. https://doi.org/10.51983/ajms-2022.11.1.3087

Tschannen, M. M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783-805. https://doi.org/10.1016/S0742-051X(01)00036-1