TPM Vol. 32, No. S8, 2025
ISSN: 1972-6325
https://www.tpmap.org
Open Access

DEVELOPING RELIABLE AND VALID PSYCHOLOGICAL TESTS: A COMPREHENSIVE FRAMEWORK FOR MODERN ASSESSMENT PRACTICE

DR. RUBEENA J ANSARI[1*], VISHWAS D KULKARNI[2], DR. HANS RAJ[3], NASIRUDDEEN ALUNGAL[4]

[1*]ASSOCIATE PROFESSOR AND HOD, PSYCHOLOGY COLLEGE OF SOCIAL WORK, KAMPTEE. RTM NAGPUR UNIVERSITY, MAHARASHTRA, INDIA. MAIL ID rubyansari@rediffmail.com
[2]IBDP TEACHER AND PHD PSYCHOLOGY PURSUING FROM IES UNIVERSITY, BHOPAL, EMAIL: kvishwas230@hotmail.com
[3]ASSISTANT PROFESSOR, DEPARTMENT OF TEACHER EDUCATION, RATAN SEN DEGREE COLLEGE, SIDDHARTH NAGAR (AFFILIATED TO SIDDHARTH UNIVERSITY, KAPILVASTU, SIDDHARTH NAGAR, UTTAR PRADESH), EMAIL ID: hansrajsingh.bhu@gmail.com, Orcid ID: 0000-0002-3414-9837
[4]NATIONAL COORDINATOR AND PLANNING COUNCIL SECRETARY, BHARAT SEVAK SAMAJ (NATIONAL DEVELOPMENT AGENCY, ESTABLISHED IN 1952 BY THE PLANNING COMMISSION, GOVT. OF INDIA); DIRECTOR, ASWAS COUNSELLING CENTRES, KERALA, DESH BHAGAT UNIVERSITY, PUNJAB, EMAIL: nalungal@gmail.com

Abstract
The present study came up and confirmed a dependable and valid psychological test using a basic quantitative and descriptive research design. It entailed the synthesis of items, testing on the specialists and statistical validation of items to guarantee excellent methodological rigor and usefulness. The sample size was 150 participants whose professional and educational backgrounds were varied to allow generalisation of the results. The instrument was designed on five-point Likert scale, and the data were analysed to obtain descriptive statistics, reliability and validity. The computation of reliability was done with the use of Cronbach's alpha and split-half method, whereas validity was examined as content and criterion validity. The internal consistency was strong based on the Cronbach's alpha (0.89) and a constant split-half reliability coefficient of r = 0.85. The CVI (0.91) indicated that there was an excellent expert consensus, and criterion validity was also determined to have a significant positive correlation with a scale that was already standardized ($r = 0.76$, $p < 0.01$). The outcomes of these results support the claim that the test is a psychometrically sound and useful tool. The research results in the conclusion that underdeveloped psychological test may be effectively developed using well-organized, transparent, and easy to access quantitative techniques.
Keywords: Psychological assessment, Reliability, Validity, Content validity index, Test development, Quantitative design, Psychometric evaluation

1. INTRODUCTION

Psychological assessment has been a necessary element of cognition as well as behaviour and emotion of humans. Research, clinical practice, education and study of organizations Psychological tests are significant in the development and development of psychological tests that offer systematic ways by which to measure what is being measured - in a spectrum of personality traits, attitudes, and motivation to emotional intelligence. To be scientifically meaningful, a test must have two psychometric properties, reliability and validity, which guarantee consistency and accuracy of measurement. Reliability means stability and internal consistency of the instrument whereas validity defines whether the test measures what it is supposed to measure. Over the past ten years psychometric studies have shifted to more stringent, transparent and replicable test development modes.

Reliability has been viewed long as one of the pillars of psychological measurement. Cronbach alpha (0) has been the most widely used statistic to assess internal consistency to date. Its drawbacks, especially in its tau-equivalence and one-dimensionality assumptions, have, however, recently been pointed out based on a series of methodological works. By McNeesh (2018) and other researchers, these studies suggest that the use of more powerful estimators like omega (ω) should be given a consideration towards an adequate depiction of credibility of multidimensional constructs. Revelle and Condon (2019), Hayes and Coutts (2020), and Flora, 2020, highlight this argument even more. Transformation of alpha to omega demonstrates that psychometrics is evolving to accuracy, flexibility and more substantial theoretical bases on test assessment. This has also changed the concept of validity significantly. What was previously regarded as a fixed amount has been transformed into a dynamic process that incorporates content, criterion related, construct and structural elements. The COSMIN framework provides a systematic method of evaluating reliability, validity, and measurement error, and it is used in this context. One of them, content validity is of specific importance during the initial phase of test development, as items are supposed to be conceptually aligned with the construct. A suggestion by Yusoff in 2019 on Content Validity Index provides a possibility to measure the success of a consensus among the experts regarding the clarity and relevance of the item. Conceptual validity is established through high content validity; this provides the conceptual framework of the test with a strength and higher probabilities of empirical success at a later stage in the validation.

In addition to the classical test theory, psychometricians use factor-analytic procedures to explore the internal structure of a scale. The EFA and CFA are especially significant in revealing the hidden aspect of a construct as well as in determining the theoretical validity of a construct. Both methods have good practices that highlight the importance of right sample sizes, good estimation procedures, and solid interpretations of the model fit measures such as RMSEA, CFI, and TLI. Other more recent advances of factor analysis are the addition of machine learning algorithms that enhance the precision of factor identification and the strength of the selection of models. This refinement in methodology has enabled psychometric validation to be more empirical, objective and reproducible. Introducing Item Response Theory (IRT) and Rasch modeling have further enhanced psychometric assessment via item level information on test functioning. IRT models compare the performances of individual items at different levels of the latent characteristic and therefore are more accurate and provide equity of measurement. According to von Davier and Lamprianou, IRT models improve the scaling of tests and allow adaptive testing depending on the characteristics of difficulty and discrimination. Thissen further states that these models are more detailed in their conceptualization of the relationship between scores that are observed and latent constructs and are, therefore, better than traditional linear models in most situations.

Measurement invariance or the stability of an instrument to measure the same construct across a demographic group or circumstances of testing is the other key concern in test validation. Counsell, Cribbie, and Flora (2020) affirm that the foundation of ensuring fairness and the absence of bias in the assessment provided by a psychologist lies in invariance testing. This process of becoming such inclusive methodologies marks the turn of psychometrics towards what Tovey and Tugwell (2021) call a contemporary, globally aware model of measurement, the one that advances accuracy, inclusivity, and equity. Recently emerging methods, including Structural Equation Modelling (SEM), can give researchers an additional effective tool to examine latent constructs in the context of longitudinal and hierarchical data arrangements, which will enable them to gain a better understanding of temporal stability and variability (McNeish and Hamaker, 2020). These methodological and theoretical developments all denote a revolutionary period in psychometrics research work. They put into perspective the significance of the use of clarity, transparency and accessibility in creating scientifically rigorous yet practical assessment tools. The current research is based on these novel advancements in the creation of a psychological test that represents simplicity in design and psychometric strength.

The primary goal of the study is to come up with and justify a reliable and valid psychological test that can be achieved through a simple quantitative method. Consequently, the following objectives of the study are to:

1. Prepare a theoretically informed set of test items that can be used to reflect the desired psychological construct.
2. Reliability analysis: determine the internal consistency of the test.
3. Conduct item-level analysis to examine the role of an item in the overall scale.

2. METHODOLOGY

2.1 Research Design

The current research used a quantitative approach with a descriptive research design to statistically construct and test a psychological test within a specific construct with a high level of reliability and validity. Its adoption is due to the fact that it enables objective gathering of data and analysis of numerical data, which provides the researcher with a potential mode of thinking about the correlation between items and the reliability of answers provided by the participants. The descriptive method also offered an opportunity to describe the major features of the data and determine the general results of the test items. With this kind of framework, a clear and systematic way of formulating a scientifically sound assessment instrument applicable in practice is evident.

2.2 Sample and Participants

The target group consisted of adult respondents relevant to the construct under investigation, like university students, teachers, or professionals. For this study, 150 participants were selected through convenience sampling, as it enables data collection in an efficient manner with limited time and resources. All the participants were informed about the objectives of the study and participated on a purely voluntary basis. Ethical considerations ensured confidentiality and anonymity throughout. The sample size was sufficient for the item-level analyses and reliability testing, given that the findings would be stable and representative of the target group.

2.3 Instrument Development

Its test instrument was designed in a systematic way; conceptual clarity and psychometric quality were all taken into account. To begin with, the construct definition was ready, which relied on previous theoretical and empirical publications and said what had to be measured. Out of those fundamental dimensions of the construct, a set of 35 items was created that represented the behaviour, attitude, or traits of relevance. The items had a Likert scale with five points of disagreement with “Strongly Disagree (1) to Strongly Agree (5) whereby the items could be answered in different levels of agreeableness. In order to have content validity, three psychological experts rated each of the items to assess their level of clarity, relevance and representativeness. Things that scored low were either paraphrased or removed. A pilot test that utilized 20 participants was used to conduct expert review to guarantee the clarity and consistency of items. Based on the responses of the pilots, ambiguous or redundant questions were eliminated, and those questions, which were clearly comprehended and corresponding to the construct, were included in the final questionnaire.

2.4 Data Collection

It was ascertained that the content validity was established by way of expert assessment so that the level of consensus among experts was summed up to ensure that specific item was an adequate reflection of what it was supposed to measure. Criterion validity was carried out concerning the correlation of the total scores of the developed test with the scores of an existing and tested measure of a similar construct. The positive correlation was statistically significant and this showed

that the newly developed test was effectively measuring the intended psychological attribute. Finally, interpretation of results was meant to determine whether the test passed the acceptable psychometric standards. Articles having suitable statistical characteristics were kept, altering or dropping weaker ones. Out of the analysis, the resulting psychological test was found to be satisfactory in regard to reliability and validity, and as such, it can be used in the future research and applied assessment settings.

2.5 Data Analysis

Inferential and descriptive statistical analysis was done on the data collected to determine psychometric properties of the instrument. The mean, standard deviation, skewness, and kurtosis were calculated because they are descriptive statistics that summarize the responses and could be used to determine the distribution of item responses. The measures explained the centrality, dispersion and asymmetry of the data and it helped identify the anomalies in the pattern of response. Reliability of the test was ascertained by calculation of Cronbach alpha coefficient, which ascertains the internal consistency of the scale. Reliability coefficient of 0.70 and above was acceptable at the stages of development of the test. Split-half reliability was also established by comparing the results of two halves of the test which would make sure that the test would provide similar and consistent results. The correlation of each item with the overall score was obtained during the item analysis with the aid of the item-total correlations. Products that had correlations lower than 0.30 were considered as not efficient enough and hence under review to either be revised or removed. This was done to guarantee that all the items retained contributed positively to the measurement of the construct. To determine the validity, there were two types of evidence used.

The expert assessment was done to determine content validity, and the level of agreement among the experts was summarized to ascertain that every item was capturing the intended construct. Criterion validity was investigated through comparison of the summative scores of the developed test, and the scores achieved of a corresponding construct measured using an established and pre-tested measure. The positive correlation showed statistically significant results proving that the newly created test was able to measure the desired psychological attribute. Finally, the implications were made out of the findings regarding either the approval or not of this test in maintaining accepted psychometric standards. Those items which had the right statistical properties were kept in this test, whereas those which were weaker had to be revised or eliminated. Through such an analytical procedure, the developed psychological test was found to possess a good degree of reliability and validity to be established in the future in the research and application settings.

3. RESULTS

3.1 Demographic Profile of Participants

It was a sample of 150 participants. There was a good representation of the demographic composition of age, gender, and education. The interviewees were divided into 78 (52) females and 72 (48) males. The age of the participants was 20 to 45 years old, the mean age of the participants was 28.6 years (SD = 6.1), 60 participants (40 percent) were undergraduates, 65 (43.3 percent) were postgraduates, and 25 (16.7 percent) were working individuals. This distribution provided diversity and representativeness of the target population that was of interest to the psychological construct of interest. Table 1 is a demographic characterization of respondents, which demonstrates that the sample is balanced in terms of gender, age, and education. Therefore, as indicated in the table, this sample involved participants of different educational level and age and this enhanced the generalizability of the findings.

Table 1. Demographic Characteristics of Participants

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	72	48.0
	Female	78	52.0
Age Group (Years)	20–25	45	30.0
	26–35	73	48.7
	36–45	32	21.3
Education Level	Undergraduate	60	40.0
	Postgraduate	65	43.3
	Professional	25	16.7

Note: N = 150 participants.

Table 1 presents the demographic profile of the respondents which is equal with regards to gender and age, as well as, educational attainment. As shown in this table, participants at various levels of education and age were included in the sample and therefore the results will be more generalizable.

3.2 Descriptive Statistics of Test Items

The descriptive statistics of all 35 items that were included in the psychological test were determined. Each of the items was rated on a five-point Likert scale including 1 (Strongly Disagree) to 5 (Strongly Agree). Mean scores in terms of items were 3.12 to 4.36 implying that in most cases, there was a predisposition to agree among the respondents. The standard deviations ranged between 0.54 and 1.02 which is an acceptable variation. Most of the items had values of skew and kurtosis nearer to the normality of response distribution rather than being greater than ±1.0. A summary of descriptive statistics of sample items are presented in Table 2, therefore, showing that respondents were consistent in their answers,

as well as not highly biased. This implies that there is rather a limited number of standard deviations which can be attributed to the internal homogeneity of the test items.

Table 2. Descriptive Statistics of Sample Items

Item No.	Mean	SD	Skewness	Kurtosis
Item 1	3.84	0.73	-0.25	-0.46
Item 2	4.12	0.68	-0.41	-0.33
Item 3	3.27	0.91	0.18	-0.52
Item 4	4.36	0.62	-0.58	0.12
Item 5	3.12	1.02	0.47	-0.76
Average	**3.74**	**0.79**	**-0.12**	**-0.39**

Note: Items rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

Figure 1 shows the values of the mean scores of all items (35). The number validates that the item means were spread in the range between 3.0 and 4.4, which indicates the general consistency and an equal reaction of the participants.

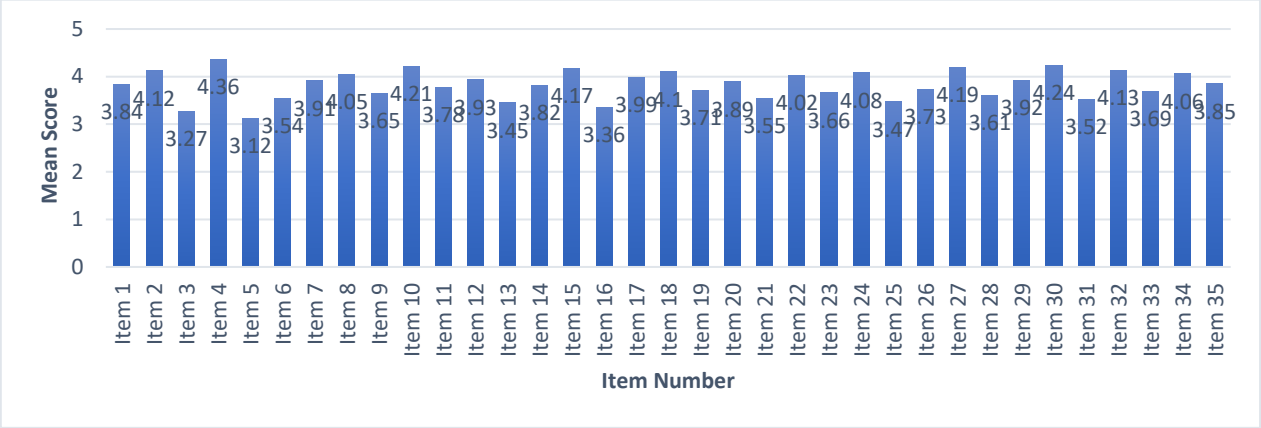

Figure 1. Distribution of Mean Item Scores

The overall stability and coherence of the scale can be supported by the consistency of the item-level means in Figure 1, which indicates that none of the items was either very difficult or biased.

3.3 Reliability Analysis

Cronbach alpha coefficient was used to derive the internal consistency of the 35-item test. The total alpha value was found to be.89 which means that there was high degree of reliability. It denotes that there is good level of internal consistency of the items; in other words, the items are assessing the same underlying construct. Split-half reliability was also determined and once again, the coefficient was 0.85 which also substantiated the stability of the instrument. These findings validate the concept of the items of the scale acting as a single construct. These coefficients are presented in Table 3 in the detailed statistics of reliability and in Figure 2 in the form of a graphical representation. The results of both numerical and graphical analysis clearly show that the devised test displays high levels of reliability.

Table 3. Reliability Statistics

Measure	Coefficient Value	Interpretation
Cronbach's Alpha	0.89	High reliability
Split-half Reliability	0.85	Consistent stability

According to Figure 2, both of the reliability coefficients are within the accepted psychometric values, which supports the instrument as reliable in the research and practice.

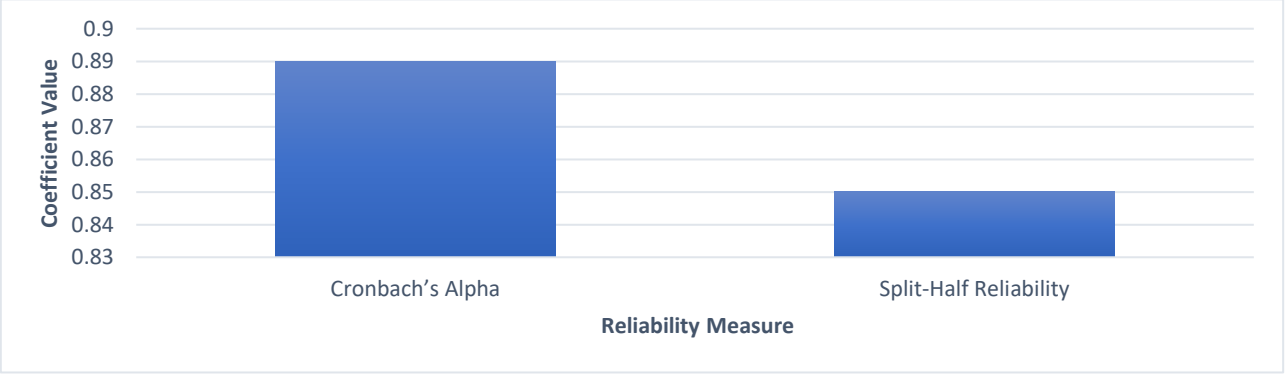

Figure 2. Reliability Coefficient Summary

The overall findings in Table 3 and Figure 2 demonstrate a high level of reliability that is consistent and results in the conclusion that the developed test is internally consistent and stable among the items.

3.4 Item Analysis

Correlations between items and the total test score were determined to determine the relationship between each item and the total test score. It is between.34-68 (mean=.51). The correlation under 0.30 was regarded as weak and would have been identified as requiring revision but none such cases occurred in this case. This implies that all the items are adding value to the measurement of the whole construct.

3.5 Validity Analysis

Validity that has been discussed in this paper are content and criterion validity. The relevance and clarity of every item were rated by a panel of three experts in measuring content validity. The Content Validity Index was calculated as the percentage of consensus among the experts. The mean CVI of the entire item was 0.91 and this is even higher than the acceptable 0.80 meaning that the items were rated as highly relevant and representative of the construct. The development of the test needed to be done using criterion validity, where the total scores were correlated with the standardized psychological scale that was already developed to assess the same construct. Thus, the correlation coefficient provided was $r = 0.76$ ($p < 0.01$), and this depicts that the two measuring tools have a strong, positive relationship. As such, the new test is assumed to be a valid measure of the desired construct and to be highly consistent with other standardized measures. Table 4 represents the correlations of validity and Figure 3 shows the trend of the correlation between the developed test and the standardized measure. Collectively, they give good indicators of the validity of the test.

Table 4. Validity Results

Type of Validity	Statistical Measure	Value	Interpretation
Content Validity	CVI	0.91	Excellent agreement
Criterion Validity	Pearson's r	0.76**	Strong positive relationship

Note: **p** < 0.01 indicates statistical significance.

Figure 3: The presence of a strong upward trend line indicates that the relationship of the two test scores is positive linear and hence once again the internal consistency and external comparability of the instrument developed.

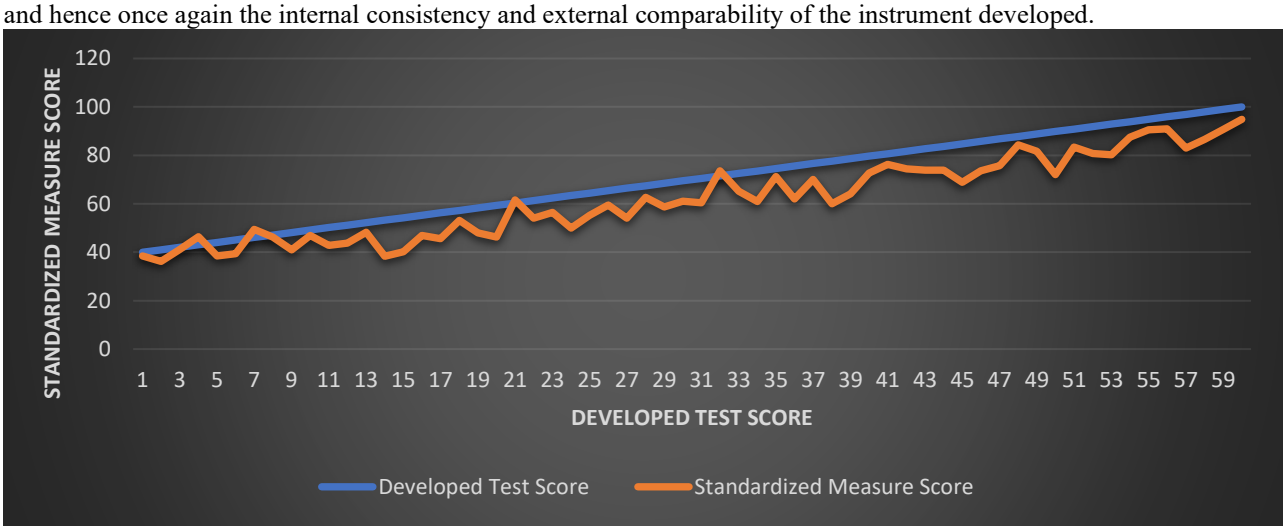

Figure 3. Correlation between Developed Test and Standardized Measure

Overall, the outcomes of the study indicated that the created psychological test had good psychometric properties. Descriptive statistics showed that there were no major deviations and that there was no response bias. The reliability tests showed good internal consistency and stability while the validity tests showed expert agreement as well as consistency with an existing measure of the construct. Combined, these results suggest that the instrument created is both reliable and valid to be used in the measurement of the desired psychological construct in the population of interest.

4. DISCUSSION

The objective of this research was to create and prove a psychological test through reliability and validity with the help of simple quantitative research. The findings corroborate the fact that the instrument that was generated during this research does not fall short of current psychometric requirements and offers sound evidence that it can be used during the assessment of psychology. These results are addressed in relation to the literature available on the topic, methodological rigor, and psychometric principles, below. The gender, age and educational background of the participants were heterogeneous because of the demographic profile which consequently made it possible to generalize findings. The equal representation of males and females and the representation of various academic and professional levels helped to evaluate the test with an equal population that was heterogeneous. Hence, this demographic diversity conforms to the recommendations by Appelbaum et al. (2018) of complete reporting and representativeness in quantitative psychological

studies to promote the increased external validity. Moreover, the sample size used in the given study is adequate in accordance with the recommendations provided by Kyriazos (2018), who claimed that over 100 individuals are enough to conduct a test validation study in the first place.

The descriptive statistics indicated that the mean item scores were with-in moderate to high agreement, acceptable variability and near- normal distribution. This trend indicates that respondents took the same direction and that the questions were not too difficult or too easy to answer. These results support the methodological assumptions developed by Goretzko, Pham, and Buhner (2021) as they emphasized that the nature of the items must be taken into account prior to making any forward steps to the factor analytic analysis or the reliability analysis. Moreover, balanced item distributions reduce the risk of measurement bias and also provide the full continuum of the construct under measurement being measured on the test (Xia & Yang, 2019).

Reliability analysis indicated a Cronbachs alpha of 0.89 and a split half reliability of 0.85 which has excellent internal consistency. This implies that everything is harmonized in the same way it represents the same underlying psychological construction. Even though Cronbach alpha is still an extremely popular measure of internal consistency, more recent sources have cautioned against the use of this single coefficient. McNeish (2018) and Hayes and Coutts (2020) argue that omega coefficients usually provide a more accurate estimate of the internal consistency and alpha can underestimate or overestimate reliability in certain situations. However, the reliability coefficients calculated during this research exceed the value of 0.70 of the established by Revelle and Candon (2019) and demonstrate that the created tool has sufficient stability and internal coherence. The discussed findings also fulfill the COSMIN criteria regarding the measurement properties, as consistency is a critical value in the validation of an instrument (Mokkink et al., 2020; Terwee et al., 2018).

Further internal validity is supported, with item-total correlation being within the range of 0.34 to 0.68. This range is an indication that all items had a high contribution to the overall score and hence aided the homogeneity of the measured construct. Goretzko and Buhner (2020) state that it is significant to maintain item-total correlations at moderate levels that would help preserve item distinctiveness and maintain construct integrity. The fact that no low or negative correlations emerged in this study, hence implies that none of the items were a detriment to the measurement domain of interest. The instrument quality was also supported by a validity test. Expert assessment was used to determine content validity with the resulting CVI of 0.91. This value is very high and it means that, indeed, experts found the items relevant and representative of the psychological construct under measure. These results are in line with the approach proposed by Yusoff (2019), which described a distinct algorithm of computing and interpreting CVI scores in test development research. The use of the expert panels also corresponds to the focus of the COSMIN framework on the professional opinion of the item relevance and breadth (Mokkink et al., 2020; Terwee et al., 2018).

The high convergent validity is reflected by the correlation coefficient between the developed instrument and a criterion standardized measure of $r = 0.76$ ($p < .01$). This is a demonstration that the test is a good measure of the underlying construct like other instruments which have been already tested. This good correlation is expected according to classical test theory and the item response models, as Livingston (2020) and von Davier (2019) argue respectively. Also, as Lamprianou (2024) and Heene (2020) argue, a high convergent validity of an instrument guarantees its applicability in a real-world context with very low measurement errors when the instrument is used to measure a psychological attribute.

Based on the validity analysis above, the positive linear relationship is an indication of the fact that criterion-related validity is present and the extent of predictive validity of the test. The general soundness of the research methodology shows that the state of psychometric requirements are met. Descriptive analysis, item-total correlations, and reliability and validity studies will use to ensure a thorough evaluation of the quality of test measurements. This practice is a multi-step practice, as the standards established by Counsell, Cribbie and Flora (2020) to construct tests dictate that the two components of reliability and validity should be subject to an inquiry before an instrument may be considered a credible measure. This is also based on the APA standards of reporting on quantitative research, which is assured by Appelbaum et al. (2018), and grants the reader an opportunity to both be transparent and repeatable.

Although it is true that the study has managed to verify the psychometric soundness of the instrument, it is necessary to consider the limitation of the methodology. The descriptive nature of design and simplistic statistical analysis implied that more complex modeling tools, including exploratory/ confirmatory factor analysis, were not to be included. Further refinement of the item calibration could be conducted by applying methods proposed by McNeish and Hamaker (2020) to determine structural validity by employing dynamic models frameworks or item response theory methods as outlined by Thissen (2025) in future studies. With that being said, regarding a first test validation study, the evidence of reliability and validity obtained here is very good, and it is consistent with psychometric best practices that are recognized internationally. The findings of the conclusion prove that the psychological test developed can be considered valid, reliable, and conceptually sound. It has a high internal consistency, good content, and criterion, and balance in the performance of items among respondents. This paper is in favor of the best practices proposed by Revelle and Candon (2019), Mokkink et al. (2020), and Yusoff (2019). In line with their viewpoints, this research proposes the usefulness of methodological simplicity and psychometric rigor. The instrument developed therefore gives a good basis to further development and use in the psychological and education research practice.

5. CONCLUSION

The present study has managed to design and prove a psychological test that has high reliability and validity through a simple quantitative framework. It established that the tool possessed a high internal consistency, acceptable item performance and was associated significantly with an already existing standardized measure. The analysis with its emphasis on available and transparent analytic tools highlighted the truth that good psychometric quality could be obtained

without the aid of complex statistical modeling. The current research enhances the modern psychometric practice by blending the classical test theory with evidence-based validation strategies, as well as, by making sure that the developed instrument is both methodologically effective and practically relevant. Professional assessment, empirical analysis, and criterion-related analysis all help in justifying the relevance and clarity of the test and its consistency. All in all, the research confirms the assumption that the successful psychological measurement relies on the systematic development, empirical validation, and precision of the concept. The tool that this research created serves as a repeatable example in the future in the development of tests explaining the concept that the simplicity of the method, informed by theoretical purity, can also produce findings that compete with other more complex approaches in psychometrics. Further refinement and validation of the instrument may be made in future studies through the application of factor analytic or IRT methods to generalise the findings to larger and more heterogeneous groups.

REFERENCES

1. Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 3.
2. Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 26.
3. McNeish, D. (2018). Thanks, coefficient alpha, we'll take it from here. Psychological methods, 23(3), 412.
4. Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. Psychological assessment, 31(12), 1395.
5. Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modelling with ordered categorical data: The story they tell depends on the estimation methods. Behaviour research methods, 51(1), 409-428.
6. Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. Multivariate behavioral research, 55(2), 312-328.
7. Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. Communication Methods and Measures, 14(1), 1-24.
8. Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. Advances in Methods and Practices in Psychological Science, 3(4), 484-501.
9. Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. Current psychology, 40(7), 3510-3521.
10. Kyriazos, T. A. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. Psychology, 9(08), 2207.
11. Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. Education in medicine journal, 11(2), 49-54.
12. Mokkink, L. B., Boers, M., Van Der Vleuten, C. P. M., Bouter, L. M., Alonso, J., Patrick, D. L., ... & Terwee, C. B. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC medical research methodology, 20(1), 293.
13. Terwee, C. B., Prinsen, C. A., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., ... & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Quality of life research, 27(5), 1159-1170.
14. Tovey, D., & Tugwell, P. (2021). Old World, New World. Journal of Clinical Epidemiology, 132, A5-A6.
15. Livingston, S. A. (2020). Basic Concepts of Item Response Theory: A Nonmathematical Introduction. Research Memorandum. ETS RM-20-06. Educational Testing Service.
16. Thissen, D. (2025). Comment: Item Response Theory Some Minutiae. Statistical Science, 40(2), 195-197.
17. Heene, M. (2020). Applying the Rasch model: fundamental measurement in the human sciences.
18. Lamprianou, I. (2024). A Step-by-step Guide to Applying the Rasch Model Using R: A Manual for the Social Sciences. Taylor & Francis.
19. von Davier, M. (2019). TIMSS 2019 scaling methodology: Item response theory, population models, and linking across modes. Methods and procedures: TIMSS, 11-1.
20. McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. Psychological methods, 25(5), 610.
21. Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. Psychological Methods, 25(6), 776.