Iranian Journal of Psychoeducational Assessment
University of Zanjan

Spring & Summer 2025. Vol 1. Issue 1

Research Paper

Cross-Cultural Psychometrics of Measurement Instruments: Strategies for Evaluation and Research Reporting

Majid Yousefi Afrashteh[1*]

1. Associate Professor, Department of Psychology, Faculty of Humanities, University of Zanjan, Zanjan, Iran

Article info:
Received: 28.05.2025
Revised: 07.06.2025
Accepted: 16.07.2025

Keywords:
cross-cultural psychmetrics, item analysis, validity, reliability

Abstract

This article presents a structured framework for conducting and reporting cross-cultural psychometric research, aimed at addressing practical challenges and enhancing methodological coherence. Cross-cultural psychometrics involves the evaluation and adaptation of psychological measurement instruments across diverse cultural contexts to ensure conceptual equivalence, validity, and reliability. The proposed framework consists of six core stages. First, instrument selection based on theoretical foundations ensures alignment between the instrument's conceptual model and the target culture; if misaligned, the development of a new instrument is recommended. Second, cross-cultural adaptation follows Beaton et al.'s (2000) five-step process: initial translation by two bilingual translators (one informed, one blind), synthesis of translations, back-translation, expert committee review for semantic and conceptual equivalence, and cognitive debriefing or pilot testing. Third, sampling and implementation involves determining sample size using either a subject-to-variable (STV) ratio of 10:1 or a minimum of 200 participants, with ethical recruitment and confidentiality safeguards. Fourth, item analysis includes statistical indicators such as mean, standard deviation, skewness, kurtosis, item discrimination, and item-total correlation (≥0.30). Fifth, reliability evaluation employs Cronbach's alpha (≥0.70), McDonald's omega, and intraclass correlation coefficient (ICC), with ICC values interpreted from poor (0.50) to excellent (0.90). Sixth, validity assessment encompasses face and content validity (qualitative and quantitative), criterion, convergent, discriminant, factorial, and known-group validity. Confirmatory factor analysis (CFA) is recommended for factorial validity, with acceptable model fit indicated by CFI and TLI > 0.95, RMSEA < 0.06, and SRMR < 0.08. Known-group validity is supported when the instrument significantly differentiates between groups with varying levels of the underlying construct.

Use your device to scan and read the article online

Citation: Yousefi Afrashteh, M. (2025). Cross-cultural psychometrics of measurement instruments: Strategies for evaluation and research reporting. *Iranian Journal of Psychoeducational Assessment*, 1 (1), 1-23. https://doi.org/: 10.30470/ijpa.2025.726636

***Corresponding Author:** Majid Yousefi Afrashteh
Address: Department of Psychology, Faculty of Humanities, University of Zanjan, Zanjan, Iran
Email: yousefi@znu.ac.ir

Extended Abstract

Introduction

One of the most significant developments in psychometrics over recent decades has been the emergence of cross-cultural psychometric research. Cross-cultural psychometrics refers to the process of evaluating and adapting psychological measurement instruments across diverse cultural contexts to ensure conceptual equivalence, validity, and reliability, thereby enabling meaningful comparisons (Dai et al., 2024). When applying an instrument outside its original cultural setting, proper adaptation is essential. Self-report measures may be subject to bias due to factors such as social desirability, concealment, and response styles (Cruchinho, 2024), which underscores the importance of using standardized and validated instruments. Cultural adaptation offers notable advantages over developing new instruments, including reduced cost and time. This process involves a series of steps beginning with cultural adaptation based on standardized guidelines, followed by rigorous assessment of reliability and validity using accepted scientific criteria (Lacko et al., 2022). Flores-Kanter & Mosquera (2023) have criticized existing validation protocols as insufficient, calling for improved standards and procedures. The present study responds to these critiques by aiming to enhance the quality and consistency of cross-cultural psychometric research. It proposes a six-step framework derived from previous studies and expanded with additional details to guide future research, reporting, and evaluation in this field.

Instrument selection based on theoretical foundations: The initial step in evaluating and selecting instruments from different cultures involves examining their theoretical underpinnings. It is essential to determine whether the theoretical basis of the instrument is compatible with the target culture. If the selected tool lacks a relevant or culturally appropriate theoretical framework, the development of a new instrument is recommended.

Cross-cultural adaptation: According to Beaton et al. (2000), this process involves five stages: (1) initial translation by at least two bilingual translators, one informed and one blind to the process; (2) synthesis of translations with resolution of discrepancies, ideally by a third party; (3) back-translation of the synthesized version by at least two bilingual individuals; (4) expert committee review involving methodologists, subject matter experts, linguists, and translators to assess semantic, idiomatic, experiential, and conceptual equivalence; and (5) cognitive debriefing or pilot testing with a sample of respondents to refine item content.

Sampling and implementation: Sample size may be determined using the subject-to-variable (STV) ratio. This study recommends using either an STV ratio of 10:1 or a minimum of 200 participants, whichever yields the larger sample. Recruitment may be conducted through convenience sampling, personal networks, or online platforms, with strict adherence to ethical standards and confidentiality. Incentives may be used but require careful ethical consideration.

Item analysis: Statistical indicators such as mean, standard deviation, skewness, kurtosis, item discrimination index, and response distribution are recommended. A minimum item-total correlation of 0.30 with statistical significance is considered acceptable.

Reliability evaluation: Cronbach's alpha and intraclass correlation coefficient (ICC) are key indicators of test-retest reliability. Given critiques of Cronbach's alpha, McDonald's omega is recommended. When a measurement instrument is multidimensional, reporting Cronbach's alpha for the overall scale is inappropriate. However, if the factor scores can logically be aggregated to compute a total score, Cronbach's alpha may be reported for both the subscales and the overall scale. An alpha coefficient of 0.70 is generally considered the acceptable threshold for internal consistency. The use of fixed cut-off points for interpreting alpha has also been challenged. ICC values below 0.50 indicate poor reliability, 0.50–0.75 moderate, 0.75–0.90 good, and above 0.90 excellent reliability.

Validity evaluation: Face and content validity can be evaluated both qualitatively through expert consensus and quantitatively using impact score (for face validity), and CVI and CVR

indices (for content validity). Criterion validity involves correlating test scores with external benchmarks. Convergent validity is supported by high correlations with similar instruments, while discriminant validity is indicated by low correlations with unrelated constructs. Factorial validity is best assessed using confirmatory factor analysis (CFA) in cross-cultural contexts, where item-factor relationships must be statistically significant and overall model fit evaluated using fit indices. Decision-making regarding these indices is not based on criteria such as statistical significance, and various cutoff points have been reported across sources, which represents a major challenge in evaluating the validity of measurement instruments. According to Hu and Bentler (1999), the most widely cited source for goodness-of-fit indices, acceptable model fit is indicated by CFI and TLI values greater than 0.95, RMSEA less than 0.06, and SRMR less than 0.08. Known-group validity refers to the instrument's ability to distinguish between groups expected to differ on the measured construct. The most commonly used estimation method is Maximum Likelihood (ML), which relies on assumptions such as multivariate normality, linear relationships between indicators and constructs, continuous variables, and independence of observations. Multivariate normality can be assessed using various tests, including Mardia's test (1980); if the null hypothesis is confirmed, skewness and kurtosis do not significantly deviate from a normal distribution. Weighted Least Squares Mean and Variance adjusted (WLSMV) is an alternative estimation method suitable when the assumption of multivariate normality is violated. It is also more appropriate when variables are categorical (e.g., Likert-type scales) or when sample size is insufficient (Kline, 2023).

Conclusion

This article aimed to propose a structured framework for conducting and reporting cross-cultural psychometric research, offering practical recommendations for decision-making challenges based on widely cited and foundational sources. Cross-cultural psychometrics in this study was outlined through six key stages: instrument selection based on theoretical foundations, cross-cultural adaptation, sampling and implementation, item analysis, reliability assessment, and validity evaluation. These stages can enhance the rigor, consistency, and trustworthiness of psychometric studies and serve as a benchmark for evaluating measurement tool reports. The study acknowledges certain limitations, such as the exclusion of item response theory and some aspects of classical test theory, and recommends that future researchers address these areas in their own work.

Ethical Considerations

This was a review study and did not involve human participants. Nevertheless, the author adhered to ethical principles, including academic integrity and proper citation of credible sources.

Funding

All expenses related to the research and manuscript preparation were fully covered by the author. This study did not receive any external funding or financial support.

Author Contributions

All activities including literature search and evaluation, translation, manuscript development, and writing were solely undertaken by Majid Yousefi Afrashteh.

Conflict of Interest

The author declares that there is no conflict of interest related to this study.

دو فصلنامه علمی سنجش روانی تربیتی
دانشگاه زنجان

مقاله پژوهشی

روانسنجی بین‌فرهنگی ابزارهای اندازه‌گیری: راهبردهایی برای ارزیابی و گزارش پژوهشی

مجید یوسفی افراشته[1*]

۱. دانشیار گروه روانشناسی، دانشکده علوم انسانی، دانشگاه زنجان، زنجان، ایران

اطلاعات مقاله:
تاریخ دریافت: ۱۴۰۴/۰۳/۰۷
تاریخ داوری: ۱۴۰۴/۰۳/۱۷
تاریخ پذیرش: ۱۴۰۴/۰۴/۲۵

واژه‌های کلیدی:
روانسنجی بین‌فرهنگی، تحلیل سؤال، روایی، اعتبار

چکیده

این مقاله چارچوبی ساختارمند برای انجام و گزارش پژوهش‌های روانسنجی بین‌فرهنگی ارائه می‌دهد که هدف آن پاسخ به چالش‌های عملی و ارتقای انسجام روش‌شناختی است. روانسنجی بین‌فرهنگی به ارزیابی و انطباق ابزارهای اندازه‌گیری روان‌شناختی در زمینه‌های فرهنگی مختلف اشاره دارد تا معادل‌سازی مفهومی، روایی و اعتبار آنها تضمین شود چارچوب پیشنهادی در این مطالعه شامل شش مرحله اصلی است. نخست، انتخاب ابزار بر اساس مبانی نظری که تضمین می‌کند مبنای مفهومی ابزار با فرهنگ مقصد سازگار است؛ در صورت عدم انطباق، ساخت ابزار جدید توصیه می‌شود. دوم، انطباق بین‌فرهنگی که طبق فرآیند پنج‌مرحله‌ای بیتن و همکاران (۲۰۰۰) انجام می‌شود: ترجمه اولیه توسط دو مترجم دوزبانه (یکی مطلع و دیگری بی‌اطلاع)، تلفیق ترجمه‌ها، ترجمه معکوس، بررسی توسط کمیته متخصصان برای معادل‌سازی معنایی و مفهومی و اجرای مقدماتی یا مصاحبه گروهی. سوم، نمونه‌گیری و اجرا که شامل تعیین حجم نمونه بر اساس نسبت آزمودنی به متغیر (STV) برابر با ۱۰:۱ یا حداقل ۲۰۰ نفر است، با رعایت اصول اخلاقی و محرمانگی در جذب مشارکت‌کنندگان. چهارم، تحلیل سؤال‌ها با استفاده از شاخص‌های آماری مانند میانگین، انحراف معیار، کجی، کشیدگی، شاخص تمایز و همبستگی سؤال با نمره کل (حداقل ۰/۳). پنجم، ارزیابی اعتبار با استفاده از آلفای کرونباخ (حداقل ۰/۷)، ضریب اومگای مک‌دونالد و ضریب همبستگی درون‌طبقه‌ای (ICC) که از ضعیف (۰/۵) تا قوی (۰/۹) تفسیر می‌شود. ششم، ارزیابی روایی شامل روایی صوری و محتوایی (کیفی و کمی)، ملاکی، همگرا، واگرا، عاملی و گروه‌های شناخته‌شده است. برای روایی عاملی، تحلیل عاملی تأییدی (CFA) توصیه می‌شود. برازش قابل قبول مدل با مقادیر CFI و TLI بیشتر از ۰/۹۵، RMSEA کمتر از ۰/۰۶، و SRMR کمتر از ۰/۰۸ مشخص می‌شود. روایی گروه‌های شناخته شده بر اساس تفاوت معنیدار نمره‌های حاصل از ابزار اندازه‌گیری در بین گروه‌هایی که به میزان مختلفی از سازه زیربنایی برخوردار هستند.

استناد: یوسفی افراشته، م. (۱۴۰۴). روانسنجی بین‌فرهنگی ابزارهای اندازه‌گیری: راهبردهایی برای ارزیابی و گزارش پژوهشی. *سنجش روانی تربیتی*، ۱(۱)، ۱-۲۳.

https://doi.org/:10.30470/ijpa.2025.726636

* **نویسنده مسئول:** مجید یوسفی افراشته
نشانی: گروه روانشناسی، دانشکده علوم انسانی، گروه روانشناسی، دانشگاه زنجان، زنجان، ایران
پست الکترونیکی: yousefi@znu.ac.ir

مقدمه

روانسنجی یک حوزه تخصصی بین رشته‌ای است که توسط فور[1] (۲۰۲۱) به عنوان علم طراحی، ارزیابی و بهبود آزمون‌های روان‌شناختی و ابزارهای اندازه‌گیری تعریف شده است. تمرکز این علم بر ویژگی‌های ابزارهای سنجش مانند اعتبار[2] و روایی[3] است، و بررسی می‌کند که چگونه این ابزارها برای اندازه‌گیری سازه‌های روان‌شناختی مانند هوش، ویژگی‌های شخصیتی، و نگرش‌های انسانی مورد استفاده قرار می‌گیرند. در واقع سازه‌ها صفات پنهان روانی هستند که به طور مستقیم قابل اندازه‌گیری نیستند و صرفاً با شناسایی و اندازه‌گیری نشانگرهایی می‌توان به طور غیرمستقیم به آنها نمره اختصاص داد. البته علم روانسنجی یک حوزه بین رشته‌ای است که در حوزه‌های گوناگون علوم رفتاری از جمله سلامت، مدیریت، علوم تربیتی و علوم اعصاب شناختی به طور گسترده استفاده می‌شود (سوان[4] و همکاران، ۲۰۲۱؛ بورسبوم[5]، ۲۰۲۲ و ایسورانو[6] و همکاران، ۲۰۲۳).

روانسنجی در دهه‌های اخیر به‌عنوان یکی از مؤثرترین و پراستنادترین شاخه‌های تخصصی روان‌شناسی مطرح شده است؛ به‌گونه‌ای که بسیاری از صاحب‌نظران آن را یکی از محورهای اصلی تولید دانش در این حوزه می‌دانند (بورسبوم و ویجسن[7]، ۲۰۱۷ و ویجسن[8] و همکاران، ۲۰۲۱). اهمیت روانسنجی در پژوهش‌های روان‌شناختی نه‌فقط به دلیل نقش مستقیم آن در طراحی ابزارهای سنجش، بلکه همچنین به خاطر جایگاه کلیدی‌اش در تفسیر داده‌ها و نتیجه‌گیری‌های مبتنی بر شواهد است. امروزه تقریباً تمامی مطالعات تجربی منتشرشده به نحوی از اصول روانسنجی بهره برده‌اند، چه در قالب استفاده از ابزارهای معتبر و چه از طریق تحلیل‌های آماری مبتنی بر مدل‌های روانسنجی (جونز و تیسن[9]، ۲۰۰۶ و فور، ۲۰۲۱ و کالکبرنر[10]، ۲۰۲۱). البته تأثیر روانسنجی محدود به حوزه‌ی پژوهش‌های تجربی نیست، بلکه در ساختار تصمیم‌گیری‌های علمی و سیاست‌گذاری‌های مربوط به نشر علمی نیز جایگاهی مهمی پیدا کرده است. بسیاری از نشریات علمی معتبر، به‌ویژه در زمینه‌های روان‌شناسی، تعلیم و تربیت، سلامت و علوم اجتماعی، ارزیابی اولیه مقالات را منوط به رعایت معیارهای دقیق روانسنجی مانند روایی و پایایی ابزارهای سنجش مورد استفاده در مطالعات کرده‌اند. به‌عبارت دیگر، استانداردهای روانسنجی به نوعی به قواعد بنیادین در نظام ارزیابی علمی بدل شده‌اند که رعایت آنها شرط اساسی برای پذیرش و انتشار پژوهش محسوب می‌شود (ایچ[11]، ۲۰۱۴ و ترافیمو و مارکس[12]، ۲۰۱۵ و پارادس و کری[13]، ۲۰۲۴).

یکی از مهمترین فعالیت‌های روانسنجی در دهه‌های گذشته روانسنجی بین فرهنگی[14] بوده است. روانسنجی بین‌فرهنگی به فرآیند ارزیابی و انطباق ابزارهای اندازه‌گیری روان‌شناختی در زمینه‌های فرهنگی مختلف اشاره دارد. هدف اصلی آن، اطمینان از این است که یک ابزار سنجش در فرهنگ‌های گوناگون، مفهوم، اعتبار و روایی یکسانی داشته باشد و نتایج حاصل از آن قابل مقایسه و تفسیر باشند (کانث[15] و همکاران، ۲۰۲۴ و دای[16] و همکاران، ۲۰۲۴). برای استفاده از هر ابزار سنجش در فرهنگی متفاوت از فرهنگ اصلی آن، لازم است ابزار اندازه‌گیری به‌درستی انطباق داده شود. باید توجه داشت که ابزارهای خودگزارشی ممکن است در معرض تحریف قرار گیرند؛ عواملی مانند تمایل به پاسخ‌های اجتماعی مطلوب، پنهان‌کاری، و سبک پاسخ‌دهی می‌توانند بر نتایج تأثیر بگذارند (کروچینو[17]، ۲۰۲۴). از این‌رو، تأکید زیادی بر استفاده از ابزارهای پژوهشی استاندارد و اعتبارسنجی‌شده برای سنجش پاسخ‌ها وجود دارد. علاوه بر این، انطباق فرهنگی یک ابزار مزایای زیادی نسبت به توسعه ابزار جدید دارد؛ از جمله کاهش هزینه‌ها و زمان صرف‌شده برای طراحی. فرآیند انطباق فرهنگی و اعتبارسنجی روانسنجی شامل مجموعه‌ای از مراحل است که با انطباق فرهنگی طبق دستورالعمل‌های استاندارد آغاز می‌شود و سپس با ارزیابی انواع مختلف روایی و اعتبار با استفاده از معیارهای علمی پذیرفته‌شده ادامه می‌یابد (لاکو[18] و همکاران، ۲۰۲۲).

با وجود گسترش اثرگذاری روانسنجی در حوزه پژوهشی، انتقادات مختلفی به آن شده است. سالزبرگر[19] (۲۰۱۳) در جمع بندی این انتقادات آنها را به دو دسته بیرونی و درونی تقسیم کرده است. انتقادات بیرونی به شکل بدبینانه‌ای حول این سؤال تمرکز یافته‌اند که که آیا روانسنجی در طول دهه‌ها فعالیت خود اصلاً سودمند بوده است یا خیر. صاحب‌نظران مختلفی در این بحث ماهوی روانسنجی مشارکت کرده و به ابهامات پاسخ داده‌اند می‌کنند (تیلور و ردفورد[20]، ۱۹۸۶؛ میشل[21]، ۲۰۰۸؛ اوهر[22]، ۲۰۲۱ و ویجسن و همکاران، ۲۰۲۲) و پاسخ به آن‌ها خارج موضوع این مقاله است. اما عمده انتقادات درونی بر بهبود ابزارها و فرایند روانسنجی تمرکز داشته‌اند. این گروه تاکید می‌کنند پژوهش‌هایی که به ساخت و اعتباریابی ابزارهای اندازه‌گیری می‌پردازند تفاوت‌های زیادی دارند که باعث سردرگمی بهره‌برداران و همچنین دامن زدن به بحران‌هایی مثل عدم تکرارپذیری (استیونز[23]، ۲۰۱۷)، فقدان پروتکل‌های پیش ثبت (اسپیتزر و مولر[24]، ۲۰۲۳) و شائبه دست‌کاری آماری (استفان، شانبرود[25]، ۲۰۲۳) می‌شوند. این انتقادات درباره مبانی نظری و سودمندی ابزارهای

1 Furr
2 Reliability
3 Validity
4 Swan
5 Borsboom
6 Isvoranu
7 Borsboom & Wijsen
8 Wijsen
9 Jones & Thissen
10 Kalkbrenner
11 Eich
12 Trafimow & Marks
13 Paredes & Carré
14 Cross-cultural psychometrics
15 Kanth
16 Dai
17 Cruchinho
18 Lacko
19 Salzberger
20 Taylor Radford
21 Michell
22 Uher
23 Stevens
24 Spitzer & Mueller
25 Stefan & Schönbrodt

روانسنجی توافق دارند اما روش‌ها و شیوه‌نامه‌های رواسازی ابزارها را کافی نمی‌دانند و بر لزوم بهبود در استانداردها و روش‌های اجرایی تأکید می‌کنند (فلورس-کانتر و موسکرا[1]، ۲۰۲۳).

پژوهش حاضر در جهت پاسخ به انتقادات گروه دوم و تقویت کیفیت و هماهنگ‌سازی پژوهش‌های روانسنجی بین فرهنگی تنظیم شده است. منظور از ابزار اندازه‌گیری در پژوهش حاضر شامل طیف گسترده‌ای از روش‌های سنجش کمی شامل روانسنجی، مقیاس، سیاهه و آزمون است که در طول مقاله با عنوان کلی ابزار اندازه‌گیری معرفی شده‌اند و می‌تواند شامل هر یک از مصادیق باشد. این ابزارها ممکن است به سنجش شخصیت، توانایی یا نگرش‌های افراد بپردازند. انتظار می‌رود نتایج این پژوهش در اجرای پژوهش‌ها، گزارش‌ها و ارزیابی علمی مقالات منتشر شده در حوزه روانسنجی مفید باشد.

مروری بر پیشینه پژوهش

طی سال‌های گذشته مراحل مختلفی برای ساخت ابزار اندازه‌گیری از طرف پژوهشگران و انجمن‌های علمی پیشنهاد شده است. برای مثال کراکر و آلجینا[2] (۱۹۸۶) یازده مرحله، فور (۲۰۱۱) پنج مرحله، استرینر[3] و همکاران (۲۰۱۶) دویلز[4] (۲۰۱۷) هفت مرحله و ایروین و هوگس[5] (۲۰۱۸) نه مرحله را پیشنهاد کرده‌اند. کریازوس و استالیکاس[6] (۲۰۱۸) با ترکیب مدل‌های مختلف الگوی یکپارچه شامل پنج فاز که هر فاز شامل چند مرحله است را پیشنهاد کرده‌اند که به ترتیب عبارت‌اند از: ۱) هدف ابزار و سازه مورد اندازه‌گیری، ۲) مشخص کردن مقیاس پاسخ‌گویی، ۳) نوشتن سؤال‌ها، ۴) ارزیابی سؤال‌ها، ۵) آزمون ویژگی‌های روانسنجی ابزار. کالکبرنر (۲۰۲۱) مدل هفت مرحله‌ای جامع و مبتنی بر شواهد خود با عنوان و سرواژه MEASURE برای توسعه و اعتبارسنجی نمرات ابزارهای اندازه‌گیری معرفی کرده است. این مراحل شامل: روشن‌سازی هدف و منطق پژوهش، ایجاد چارچوب تجربی، تبیین نقشه نظری، تلفیق محتوا و ساخت مقیاس، ارزیابی متخصصان، جذب شرکت‌کنندگان و اجرا، و ارزیابی اعتبار و روایی ابزار می‌باشند. این رویکرد جامع، مسیر منطقی و ساختارمندی برای طراحی ابزارهای روانسنجی و سنجش فراهم می‌کند که در مطالعات علمی و بالینی از اهمیت بالایی برخوردار است.

علاوه بر این دستورالعمل‌های متعددی برای کمک به نویسندگانی که فرآیند توسعه ابزارها و مقیاس‌های جدید روانسنجی را گزارش می‌دهند، تدوین شده‌اند. قدیمی‌ترین و جامع‌ترینِ این دستورالعمل‌ها، استانداردهای سنجش آموزشی و روان‌شناختی[7] است که در سال ۱۹۹۹ با حمایت مشترک انجمن پژوهش آموزشی آمریکا[8]، انجمن روان‌شناسی آمریکا[9] و شورای ملی اندازه‌گیری آموزشی[10] منتشر شده است. اگرچه این اثر به‌عنوان کتاب مقدس گزارش‌نویسی در روانسنجی شناخته می‌شود، مخاطب اصلی آن سازندگان آزمون‌های پرخطر[11] هستند؛ آزمون‌هایی مانند سنجش پیشرفت تحصیلی، تعیین سطح مدرسه، یا آزمون‌های ورودی کارشناسی‌ارشد و دکتری حرفه‌ای (استرینر و کاتنر[12]، ۲۰۱۴).

دستورالعمل استانداردهای گزارش‌دهی دقت تشخیصی[13] (STARD) توسط باسیوت[14] و همکاران (۲۰۰۳) مطرح شده است و هدف آن بهبود گزارش‌ها در مطالعات مربوط به دقت تست‌های تشخیصی بوده است. این راهنما ارزیابی جنبه‌های کلیدی مطالعات و یافته‌های آنها را ممکن می‌سازد، از جمله بررسی میزان سوگیری و قابلیت کاربرد نتایج (وایت[15] و همکاران، ۲۰۲۵). همان‌طور که نام آن نشان می‌دهد، STARD بیشتر به حوزه پزشکی مربوط است و صرفاً بر آزمون‌های تشخیصی توجه دارد. این استاندارد شامل یک چک‌لیست، یک نمودار جریان، و توضیحات تفصیلی است (کوهن[16]، ۲۰۱۶). در ابتدا برای نویسندگان مجلات پزشکی طراحی شد، اما بعدها برای پوشش آزمون‌های تشخیصی با رویکرد روان‌شناختی نیز اصلاح گردید (میر[17]، ۲۰۰۳). امروزه بسیاری از مجلات پزشکی و بیوپزشکی برجسته که گزارش‌های آزمون‌های تشخیصی را منتشر می‌کنند، این دستورالعمل‌ها را پذیرفته‌اند. این دستورالعمل نیز محدودیت استفاده برای گروه خاصی از ابزارها با عنوان آزمون‌های تشخیصی دارند و قابل استفاده برای طیف وسیع ابزارهای اندازه‌گیری نیستند. آزمون‌های تشخیصی آزمون‌هایی هستند که در آنها هدف سنجش وجود یا عدم بیماری است (استرینر و کاتنر، ۲۰۱۴ و استاهل[18] و همکاران، ۲۰۲۳).

دستورالعمل گزارش‌دهی مطالعات اعتبار و توافق[19] (GRRAS) توسط کاتنر[20] و همکاران (۲۰۱۱) تهیه و منتشر شده که شامل ۱۵ معیار کلیدی هستند و باید در مقالاتی که بر مطالعات اعتباریابی آزمون‌ها تمرکز دارند، مورد توجه قرار گیرند. محورهای کلیدی این دستورالعمل عبارت‌اند از: مشخص‌کردن نوع توافق (بین ارزیاب‌ها یا درون‌ارزیاب)، توصیف ابزار یا روش اندازه‌گیری، ویژگی‌های جمعیت مورد مطالعه و ارزیاب‌ها، روش نمونه‌گیری و تعداد ارزیابی‌ها، تحلیل آماری مورد استفاده و نحوه محاسبه توافق یا قابلیت اعتماد، گزارش نتایج همراه با عدم قطعیت آماری (مثل فاصله اطمینان).

1 Flores-Kanter & Mosquera
2 Crocker & Algina
3 Streiner
4 DeVellis
5 Irwing & Hughes
6 Kyriazos & Stalikas
7 The Standards for Educational and Psychological Testing
8 American Educational Research Association
9 American Psychological Association
10 National Council on Measurement in Education
11 High-Stake
12 Streiner & Kottner
13 The Standards for Reporting of Diagnostic Accuracy
14 Bossuyt
15 White
16 Cohen
17 Meyer
18 Stahl
19 Guidelines for Reporting Reliability and Agreement Studies
20 Kottner

توصیه‌های فراهم شده در این دستورالعمل نیز بر ویژگی اعتبار ابزارهای اندازه‌گیری تمرکز دارد؛ در صورتی که ابزارسازی یا روانسنجی فرایند بسیار جامع‌تری دارد.

این دستورالعمل‌ها بیشتر زمینه محدودی را پوشش می‌دهند و فرایند جامع مربوط به روانسنجی را ارائه نمی‌کنند. علاوه بر این هیچ یک از آنها در مراحل خود اشاره‌ای به روانسنجی بین فرهنگی نکرده‌اند. در این بین دستورالعملی که توسط بیتن[1] و همکاران (۲۰۰۰) برای روانسنجی بین فرهنگی ارائه شده، رایج‌ترین و پرکاربردترین راهنما در این زمینه است. مهمترین افزوده این دستورالعمل معادل‌سازی[2] است. بیتن (۲۰۰۰) بر چهار نوع معادلسازی معنایی[3] (حفظ معنی اصلی سؤال‌ها)، اصطلاحی[4] (ترجمه صحیح اصطلاحات و زبان عامیانه)، تجربی[5] (تطابق تجربه‌های روزمره در فرهنگ هدف) و مفهومی[6] (حفظ مفهوم اصلی سؤال‌ها با توجه به تفاوت‌های فرهنگی) تاکید کرده است. او برای اطمینان از معادل‌سازی شش مرحله عملیاتی پیشنهاد کرده است: ۱)ترجمه اولیه، ۲) تلفیق ترجمه‌ها توسط نفر سوم، ۳) ترجمه معکوس نسخه تلفیقی، ۴) بررسی ترجمه‌ها در کمیته‌ای از متخصصان، ۵) پیش اجرای مقدماتی و ۶) مستندسازی و نهایی‌سازی ابزار. همان‌طور که مشخص است مراحل بیتن و همکاران (۲۰۰۰) بر تطبیق فرهنگی و معادل‌سازی تمرکز دارند و به جزئیات ارزیابی‌های فنی روانسنجی اشاره نکرده است. لنز[7] و همکاران (۲۰۱۷) شش مرحله برای تطبیق و روانسنجی بین‌فرهنگی پیشنهاد کرده‌اند که عبارت‌اند از: ترجمه اولیه، ترجمه معکوس، بررسی در هیئت کارشناسی، اجرای مقدماتی، تحلیل روانسنجی و مستندسازی و گزارش نهایی. هدف اصلی این پژوهش فراهم کردن جزئیات اجرایی برای همه مراحل ممکن ابزارسازی است. بنابراین در پژوهش حاضر تلاش شده است با تلفیق مدل‌ها و دستورالعمل‌های توصیه شده در مطالعات قبلی یک ساختار مناسب برای گزارش‌نویسی پژوهش‌هایی که ویژگی‌های روانسنجی بین‌فرهنگی ابزارها را گزارش می‌کنند فراهم آید. همچنین امید است این مقاله به آن دسته از بهره‌برداران که ابزارهای طراحی‌شده توسط دیگران را ارزیابی می‌کنند نیز کمک کند تا دریابند که آیا ابزار موردنظر برای استفاده آنها مناسب است یا خیر.

مراحل روانسنجی بین فرهنگی

در پژوهش حاضر مراحل روانسنجی بین فرهنگی با تلفیق مراحل گزارش شده توسط پژوهش‌های قبلی و تکمیل پاره‌ای از جزئیات حاصل شده است. ساختار پیشنهادی شامل شش مرحله کلی است که به ترتیب عبارت‌اند از انتخاب ابزار بر اساس مبانی نظری، انطباق بین‌فرهنگی، نمونه‌گیری و اجرا، تحلیل سؤال، ارزیابی اعتبار و ارزیابی روایی است.

انتخاب ابزار بر اساس مبانی نظری

انتظار می‌رود سازندگان ابزارها هدف ابزار خود را به‌روشنی تعریف کنند؛ به‌ویژه اینکه چه سازه‌ای را قصد دارند اندازه‌گیری کنند و چرا سنجش آن سازه از اهمیت برخوردار است (دویلز، ۲۰۱۷). در روانسنجی بین فرهنگی پژوهشگران باید با دقت و سخت‌گیری روش‌شناسی ابزار را ارزیابی کرده و روش‌های به‌کاررفته توسط توسعه‌دهندگان ابزار را با استانداردهای تجربی معتبر مقایسه کنند. اولین اقدام در ارزیابی و انتخاب ابزار از فرهنگ‌های مختلف توجه به خاستگاه نظری آن است. اینکه آیا مبنای نظری ابزار در فرهنگ مقصد قابلیت انطباق دارد یا خیر مهم است (دیمیتوف[8]، ۲۰۱۲). مارنات و رایت (۱۴۰۰) روایت می‌کنند که برخی از ابزارهای شناخته شده مانند MMPI در ورود به بسیاری از فرهنگ‌ها مشکل داشتند و شرکت‌کنندگان در مواجهه با برخی از سؤال‌ها که موضوعات جنسی یا رفتار توالت رفتن را شامل می‌شد دچار شوک فرهنگی می‌شدند. در شرایطی که مبنای نظری هدف یا عنوان کل ابزار یا برخی از خرده‌مقیاس‌ها با زمینه فرهنگی سازگار نباشد یا درباره عدالت بین فرهنگی تردید ایجاد کند انجام مطالعه‌ای برای توسعه ابزار جدید ضرورت می‌یابد (ون‌دویجر و تنزر[9]، ۲۰۰۴). در برخی موارد، طراحی یک ابزار اختصاصی برای جمعیت‌های متنوع، مناسب‌تر از اعتبارسنجی ابزارهای موجودی است که با جمعیتی متفاوت توسعه یافته‌اند. برای نمونه، اگر پژوهشگری به دنبال ابزاری برای غربالگری پریشانی روانی در میان مخاطب ایرانی باشد، ممکن است طراحی یک ابزار جدید مبتنی بر ویژگی‌های فرهنگی، کارآمدتر از اعتبارسنجی ابزار وارداتی باشد؛ چرا که اعتبار محتوایی سازه مورد سنجش می‌تواند در فرهنگ‌های مختلف تفاوت چشمگیری داشته باشد. حتی اگر ابزار موجود از نظر فنی و آماری مناسب تشخیص داده شود، ممکن است نتواند جنبه‌های منحصربه‌فرد پریشانی روانی در آن فرهنگ را به‌درستی بازتاب دهد (کالکبرنر، ۲۰۲۱). در نهایت، پژوهشگر باید نشان دهد که چگونه ابزار پیشنهادی می‌تواند شکاف موجود در ادبیات سنجش را پر کرده و به پیشرفت پژوهش و کاربردهای عملی در آینده کمک کند. این کار با مرور مبانی نظری زیربنایی و معرفی برخی از ابزارهای مشابه موجود در بافت فرهنگی هدف در بخش مقدمه مقاله باید صورت گیرد (سوان و همکاران، ۲۰۲۳).

انطباق بین‌فرهنگی

معادل‌سازی و هماهنگی محتوا و زبان یک ابزار در فرهنگ جدید انطباق خوانده شده است (بیتن، ۲۰۰۰ و عرفات[10]، ۲۰۱۶). این فرایند در برخی منابع به عنوان روایی زبانی[11] نیز معرفی شده است. روایی زبانی به فرایندی اطلاق می‌شود که طی آن اطمینان حاصل می‌گردد محتوای ابزار اندازه‌گیری–یا

[1] Beaton
[2] Equivalence
[3] Semantic
[4] Idiomatic
[5] Experiential
[6] Conceptual
[7] Lenz
[8] Dimitrov
[9] Van de Vijver & Tanzer
[10] Arafay
[11] Linguistic Validity

هر متن دیگری–نه‌تنها از نظر معنایی دقیق است، بلکه از نظر فرهنگی نیز مناسب و برای مخاطبان هدف قابل درک است. این فرایند چندمرحله‌ای به منظور تأیید این است که متن ترجمه‌شده، معنا و هدف اصلی نسخه اولیه را حفظ کرده و در عین حال تفاوت‌های فرهنگی و برداشت‌های احتمالی متفاوت را نیز مدنظر قرار می‌دهد (هرناندز[1] و همکاران، ۲۰۲۰ و سمیک[2] و همکاران، ۲۰۲۲).

ترجمه، نخستین مرحله در فرآیند انطباق است، اما اصطلاح انطباق با ترجمه تفاوت دارد؛ زیرا انطباق شامل تمام فرآیندهای مرتبط با جنبه‌های فرهنگی، اصطلاحی، زبانی و زمینه‌ای در ترجمه است. مراحل انطباق بین فرهنگی طبق بیتن (۲۰۰۰) به صورت زیر پیشنهاد می‌شود.

۱)ترجمه اولیه توسط حداقل دو مترجم که در هر دو زبان مهارت کافی دارند؛ یکی از آنها از فرآیند آگاه است و دیگری به‌صورت ناشناس عمل می‌کند؛
۲) تلفیق ترجمه‌ها با رفع اختلافات میان آنها، که بهتر است توسط فرد سومی انجام شود؛
۳) ترجمه معکوس نسخه تلفیقی توسط حداقل دو نفر که در هر دو زبان مهارت دارند؛
۴) بررسی مراحل توسط کمیته‌ای متشکل از روش‌شناسان، متخصصان موضوعی، متخصصان زبان و مترجمان (ترجمه مستقیم و معکوس)، با در نظر گرفتن معادل‌سازی معنایی، اصطلاحی، تجربی و مفهومی؛
۵) نشست با گروهی از پاسخ‌دهندگان جهت طرح و بررسی محتوای سؤال‌ها و یا اجرای مقدماتی[3] ابزار بازبینی‌شده.

به دلیل دشواری‌های تنظیم جلسه با گروهی از پاسخ‌دهندگان پژوهشگران معمولاً از اجرای مقدماتی استقبال می‌کنند. این اجرا فرصتی فراهم می‌آورد تا بازخورد شرکت‌کنندگان درباره محتوای سؤال‌ها و میزان خوانایی آنها دریافت شود. اگرچه دستورالعمل‌های مختلفی برای تعیین حجم نمونه در این اجرا وجود دارد، اما الزام به رعایت حداقل حجم نمونه تاکید نشده است. معمولاً این نمونه‌ها بین ۲۵ تا ۱۵۰ نفر در نظر گرفته می‌شوند (هرتزوک[4]، ۲۰۰۸). داده‌های حاصل از مطالعه آزمایشی باید از نظر وضوح، خوانایی سؤال‌ها و خطاهای اجرایی بررسی شوند. پژوهشگران می‌توانند تحلیل‌های مقدماتی سؤال‌ها مانند همبستگی بین سؤال‌ها و آمارهای توصیفی را به‌صورت آزمایشی محاسبه کنند. اگرچه حجم نمونه ۱۰۰ نفر یا بیشتر برای انجام این تحلیل‌ها مطلوب است (ترسی[5] و همکاران، ۲۰۲۲) اما در صورت استفاده از نمونه‌های کوچکتر نیز می‌توان این تحلیل‌ها را انجام داد، مشروط بر آنکه محدودیت‌های ناشی از حجم نمونه در تفسیر نتایج لحاظ شود. در صورت کافی بودن حجم نمونه پژوهشگران می‌توانند این داده‌ها را در تحلیل عاملی (مرحله روایی) نیز استفاده کنند. اگر شرکت‌کنندگان در مطالعه آزمایشی به مشکلاتی در محتوای سؤال‌ها یا خوانایی آنها اشاره کنند، پژوهشگران باید ابزار را بازنگری کرده و فرآیند اجرای مقدماتی را تکرار کنند.

نمونه‌گیری و اجرا

در مطالعات روانسنجی، پژوهشگران باید پیش از آغاز جمع‌آوری داده‌ها، حداقل تعداد نمونه مورد نیاز برای اجرای اصلی را به‌صورت پیشینی تعیین کنند. توصیه‌ها درباره حجم نمونه در برآورد اعتبار بسیار متغیر هستند و از ۲۰۰ نفر (گیلفورد[6]، ۱۹۵۶؛ کلاین[7]، ۲۰۱۳) تا ۱۰۰۰ نفر (فلد و آنکمن[8]، ۱۹۹۹) توصیه شده‌اند. با این حال، سیچتی[9] (۱۹۹۹ و ۲۰۰۱) استدلال می‌کند که افزایش دقت حاصل از حجم نمونه‌های بالاتر از ۵۰، در اغلب موارد ارزش تلاش اضافی را ندارد و حجم نمونه‌های کمتر از این مقدار باید به‌طور منطقی توجیه شوند. برآورد حجم نمونه در ارزیابی روایی متغیرتر نیز هست، زیرا طرح واحدی برای این نوع مطالعات وجود ندارد. مطالعات روایی معمولاً آزمون فرضیه هستند؛ برای مثال، تفاوت نمرات بین گروه‌ها، تغییر نمرات در پاسخ به مداخله، یا همبستگی (یا عدم همبستگی) با نمرات ابزار دیگر. هر یک از این سؤال‌ها پژوهشی به طراحی متفاوتی نیاز دارد و در نتیجه حجم نمونه متفاوتی را می‌طلبد. در هر مورد، حجم نمونه باید به‌صورت پیشینی توجیه شود. ویلیامنز[10] و همکاران (۲۰۲۵) به حجم حدود ۳۰ نفر که در برآورد همبستگی درون‌طبقه‌ای مرسوم است انتقاد می‌کند و آن را ناکافی می‌داند. او نمونه بیشتر از صد نفر را توصیه می‌کنند. کندی[11] (۲۰۲۵) حداقل ۱۰۰ شرکت‌کننده را برای محاسبه ضریب همبستگی درون‌طبقه‌ای و ضریب آلفای کرونباخ پیشنهاد کرده است. آریفین[12] (۲۰۲۵) یک ماشین حساب آنلاین را برای تعیین حجم نمونه در محاسبه چهار شاخص همبستگی درون‌طبقه‌ای، ضریب کاپا، آلفای کرونباخ و همبستگی پیرسون معرفی کرده است.

از آن‌جایی که تحلیل عاملی یکی از روش‌های آماری رایج در پژوهش‌های روانسنجی است و نسبت به سایر تحلیل‌ها به نمونه بیشتری نیاز دارد، بهتر است حداقل نمونه طبق آن تعیین شود (بندالوس و فینی[13]، ۲۰۱۹). به‌طور کلی، حجم نمونه بزرگ‌تر در تحلیل عاملی مطلوب‌تر است، زیرا قدرت آماری را افزایش می‌دهد؛ با این حال، اجماع روشنی در منابع علمی درباره تعیین حداقل حجم نمونه وجود ندارد (کنکتا[14] و همکاران، ۲۰۱۹؛ شرستا[15]، ۲۰۲۱). در گذشته، راهنمایی‌های حجم نمونه بر اساس معیارهای کلی ارائه می‌شدند. برای مثال، کومری و لی[16] (۱۹۹۲) پیشنهاد کردند: ۵۰ = بسیار ضعیف، ۱۰۰

[1] Hernández
[2] Semkiv
[3] Pilot
[4] Hertzog
[5] Teresi
[6] Guilford
[7] Kline
[8] Feldt & Ankenmann
[9] Guilford
[10] Williams
[11] Kennedy
[12] Arifin
[13] Bandalos & Finney
[14] Knekta
[15] Shrestha
[16] Comrey and Lee

= ضعیف، ۲۰۰ = متوسط، ۳۰۰ = خوب، ۵۰۰ = بسیار خوب، و بیش از ۱۰۰۰ = عالی است. در سال‌های اخیر، بسیاری از پژوهشگران روانسنجی حجم نمونه را بر اساس نسبت شرکت‌کنندگان به تعداد پارامترهای تخمینی یا متغیرهای تحلیل‌شده محاسبه می‌کنند؛ که به آن نسبت آزمودنی به متغیر[1] (STV) گفته می‌شود (بیورز[2] و همکاران، ۲۰۱۳). در مطالعات روانسنجی هر سؤال به عنوان یک متغیر در لحاظ می‌شود. این نسبت در منابع مختلف از ۱ به ۳ تا ۱ به ۲۰ متغیر است، اما نسبت ۱ به ۱۰ معمولاً قابل قبول تلقی می‌شود. با این حال، این نسبت برای ابزارهای کوتاه (حدود ۱۹ سؤال یا کمتر) ممکن است کافی نباشد (کالکبرنر، ۲۰۲۱)، زیرا در چنین مواردی حجم نمونه باید حداقل ۲۰۰ نفر باشد (کومری و لی، ۱۹۹۲).

برخی از روانسنجان معاصر رویکرد یک قاعده برای همه را رد می‌کنند و طبق توصیه آنها عوامل مختلفی مانند تعداد عامل‌ها، اشتراک‌پذیری سؤال‌ها و واریانس شرکت‌کنندگان بر حداقل حجم نمونه تأثیر می‌گذارد (بندالوس و فینی[3]، ۲۰۱۹ و کنکتا و همکاران، ۲۰۱۹). بر اساس توصیه‌های ترکیبی روان‌سنجان برجسته، کالکبرنر (۲۰۲۱) پیشنهاد می‌کند که توسعه‌دهندگان ابزار، حجم نمونه خود را بر اساس یکی از دو معیار زیر تعیین کنند (هر کدام که حجم بیشتری را نتیجه دهد):

الف) نسبت STV برابر با ۱ به ۱۰؛
ب) حجم نمونه حداقل ۲۰۰ نفر.
برای مثال، اگر ابزاری شامل ۵۰ سؤال باشد، حجم نمونه بر اساس نسبت STV برابر با ۵۰۰ نفر خواهد بود.

برای دستیابی به حجم نمونه مناسب در تحلیل‌های روانسنجی، راهبردهای مختلفی برای جذب شرکت‌کنندگان وجود دارد. نمونه‌گیری در مکان‌های عمومی (با دریافت مجوزهای لازم) می‌تواند راهبردی مقرون‌به‌صرفه باشد (شارون[4]، ۲۰۱۸). در مطالعات پیمایشی با جامعه‌ی دانشجویی، یکی از روش‌های جذب مشارکت‌کنندگان آن است که پژوهشگر هنگام ورود افراد به مکان‌های پرتردد مانند کتابخانه یا مرکز خدمات دانشجویی با آنها ارتباط برقرار کند. بهره‌گیری از شبکه‌های فردی (نظیر ارتباط با اساتید پیشین یا فعلی، همکاران یا کارفرمایان) برای ارسال دعوت‌نامه به شرکت در پژوهش نیز روش مؤثری محسوب می‌شود. به عنوان نمونه، پژوهشگر می‌تواند از یکی از اساتید بخواهد که ایمیلی برای کل دانشجویان یک گروه آموزشی ارسال کند. پژوهشگرانی که با سازمان‌هایی مانند دانشگاه‌ها همکاری دارند نیز ممکن است از پشتیبانی رسمی، به‌ویژه در فرآیند جمع‌آوری داده‌های دارای تأییدیه اخلاقی، برخوردار شوند. برخی مؤسسات دانشگاهی حتی ممکن است فهرست‌های کاملی از دانشجویان ثبت‌نام‌شده، شامل اطلاعات تماس مانند آدرس‌های ایمیل، در اختیار پژوهشگر قرار می‌دهند (کالکبرنر، ۲۰۲۱).

روش نمونه‌گیری در دسترس[5] از رایج‌ترین روش‌های جمع‌آوری داده‌ها در پیمایش‌ها و مطالعات روانسنجی محسوب می‌شود. این تکنیک نمونه‌گیری دارای چندین مزیت ذاتی است؛ از جمله هزینه پایین، صرف زمان کمتر، و سادگی در اجرا. با این حال، دارای معایبی نیز هست؛ مانند سوگیری نمونه، خطاهای نظام‌مند، نمایندگی ناکافی از جمعیت هدف، و عدم قابلیت تعمیم یافته‌های پژوهش. داده‌های حاصل از این روش معمولاً منعکس‌کننده پاسخ‌های افرادی هستند که به‌طور طبیعی تمایل یا فرصت شرکت در پژوهش‌های پیمایشی دارند (یعنی افرادی که علاقه‌مند به پاسخ‌گویی به نظرسنجی‌ها هستند). البته باید در نظر گرفت هماهنگی‌های لازم برای اجرای نمونه‌گیری‌های تصادفی در عمل استفاده از آن‌ها را سخت و محدود کرده است (سیمکوس[6]، ۲۰۱۶). اسکورونوک و دورر[7] (۲۰۰۹) پیشنهادهایی برای کنترل سوگیری‌ها و افزایش دقت این روش نمونه‌گیری ارائه کرده‌اند که می‌تواند احتمال معرف بودن نمونه را افزایش دهند. آنها پیشنهاد کرده‌اند پژوهشگران باید هنگام انتخاب شرکت‌کنندگان، سوگیری‌ها را کاهش دهند و با ارزیابی و کنترل نمایندگی نمونه، معرف بودن نمونه و تعمیم‌پذیری پژوهش را افزایش دهند. دوم، با توزیع پرسشنامه‌ها در زمان‌ها و مکان‌های مختلف، می‌توان تنوع نمونه را افزایش داد و نمایی مناسب از جامعه هدف به‌دست آورد. در نهایت، افزایش حجم نمونه نیز راهی مؤثر برای کنترل سوگیری و کاهش عدم قطعیت است؛ پژوهشگران با استفاده از نمونه‌های بزرگ‌تر می‌توانند داده‌های متنوع‌تری جمع‌آوری کنند. البته در صورت امکان استفاده از روش‌های نمونه‌گیری دقیق‌تر مانند نمونه‌گیری تصادفی، قابلیت تعمیم نتایج را افزایش می‌دهد.

پژوهشگران گاهی برای افزایش مشارکت در پژوهش‌های پیمایشی، مشوق‌های کوچکی مانند کارت هدیه الکترونیکی، خودکار یا بسته‌ای از شکلات به همه شرکت‌کنندگان ارائه می‌دهند، یا امکان شرکت در قرعه‌کشی برای دریافت جایزه را فراهم می‌سازند. با این حال، ارائه مشوق‌ها با ملاحظات اخلاقی متعددی همراه است (سینگر و بوسارت[8]، ۲۰۰۶ و نیل[9]، ۲۰۲۰). برای مثال، مشوق‌ها نباید تأثیر نامناسب یا اجبارآمیز داشته باشند؛ از جمله نباید شامل پرداخت‌های مالی بیش از حد باشند. آنچه به‌عنوان مشوق نامناسب یا افراطی تلقی می‌شود، بسته به زمینه پژوهش متفاوت است؛ بنابراین، پژوهشگران باید با تیم پژوهشی، کمیته اخلاق مؤسسه، و منابع علمی موجود مشورت کنند تا مشوق مناسب برای مطالعه خاص خود را تعیین نمایند. برای مرور کامل ملاحظات عملی و اخلاقی در زمینه مشوق‌ها، به سینگل و بوسارت (۲۰۰۶) مراجعه شود.

1 subjects-to-variables ratio
2 Beavers
3 Bandalos & Finney
4 Sharon
5 Convenience Sampling
6 Simkus
7 Skowronek & Duerr
8 Singer & Bossarte
9 Neal

پژوهش‌های مبتنی بر پرسشنامه‌ی الکترونیکی به‌طور فزاینده‌ای در حال گسترش هستند و پلتفرم‌های آنلاین نظرسنجی محبوبیت چشمگیری یافته‌اند. این پلتفرم‌ها امکانات کاربرپسند متعددی برای طراحی سؤال‌ها فراهم می‌کنند؛ از جمله ماتریس‌های ساخت مقیاس لیکرت، گزینه‌های کشویی برای مقیاس‌های بصری، پاسخ‌های نوشتاری، گزینه‌های چندگزینه‌ای و موارد دیگر (ماسلوسکایا[1]، ۲۰۲۲). اغلب این پلتفرم‌ها لینک‌های الکترونیکی ناشناس تولید می‌کنند که می‌توان آنها را در شبکه‌های مجازی برای افراد گروه هدف ارسال کرد یا در وب‌سایت‌ها منتشر نمود. پلتفرم‌های نظرسنجی الکترونیکی همچنین نیاز به ورود دستی داده‌ها را از میان برداشته‌اند، زیرا داده‌ها به‌صورت مستقیم در قالب فایل‌های SPSS یا Excel قابل دانلود هستند (کالکبرنر، ۲۰۲۱).

اگرچه در مطالعات مربوط به ساخت مقیاس‌ها معمولاً مسائل اخلاقی چندانی مطرح نمی‌شود، اما همچنان ملاحظاتی نظیر رضایت آگاهانه، احتمال فریب‌دادن شرکت‌کنندگان، حفظ محرمانگی اطلاعات، و در مواردی که سازه‌هایی مانند افسردگی مورد ارزیابی قرار می‌گیرند، وظیفه هشداردهی وجود دارد (استرینر، ۲۰۱۳). بنابراین، دریافت تأییدیه از کمیته اخلاق سازمان مربوطه (مانند دانشگاه، بیمارستان یا مدرسه) ضروری است و باید در مقاله به‌صراحت ذکر شود. طبق اساس‌نامه اخلاق در پژوهش‌های زیست پزشکی برای شرکت‌کنندگان کمتر از ۱۶ سال رضایت والدین لازم است (دبیرخانه کارگروه وزارتی اخلاق در پژوهش وزارت بهداشت، ۱۴۰۱).

تحلیل سؤال

محتوای ارائه‌شده در ابتدای یافته‌ها بسته به هدف مطالعه ممکن است تا حدی متفاوت باشد، اما برخی عناصر کلیدی همواره ثابت‌اند. نخستین بخش، توصیف دقیق محیط مطالعه و ویژگی‌های نمونه است. این توصیف باید اطلاعات کافی درباره شرکت‌کنندگان ارائه دهد تا خواننده بتواند تصویری روشن از جمعیت مورد بررسی داشته باشد؛ از جمله سن، نسبت جنسیتی، سطح تحصیلات، محل سکونت و سایر مشخصات مرتبط. در صورتی که انتخاب شرکت‌کنندگان بر اساس شرایط پزشکی یا روان‌شناختی خاصی صورت گرفته باشد، باید عوامل مرتبط با آن اختلال نیز گزارش شوند؛ برای مثال، در مطالعه‌ای با بیماران مبتلا به میگرن، اطلاعاتی مانند سن شروع بیماری، دفعات و مدت زمان سردردها، و داروهای مصرفی اهمیت دارد. اطلاعات جمعیت‌شناختی هم در ابتدای یافته‌ها و هم در بخش جامعه و نمونه می‌توانند گزارش شوند. این اطلاعات به خوانندگان درک روشنی از شرکت‌کنندگان ارائه می‌دهد که در برداشت درست از یافته‌ها مفید خواهد بود. بنابراین اخذ و گزارش اطلاعات جمعیت شناختی صرف نظر از اینکه در تحلیل‌های استنباطی استفاده شوند یا خیر ضروری و مفید هستند. در بسیاری از موارد، گزارش نرخ پاسخ‌دهی نیز ضروری است. البته در شرایطی که از کلاس‌های دانشگاهی یا گروه‌های مشخص استفاده می‌شود، این الزام وجود ندارد. همچنین، در مواردی که جذب شرکت‌کننده از طریق فهرست‌های اینترنتی انجام می‌شود، تعیین تعداد افراد واجد شرایط ممکن نیست. با این حال، زمانی که تعداد شرکت‌کنندگان بالقوه مشخص است و امکان عدم مشارکت وجود دارد، ارائه این اطلاعات برای بررسی سوگیری‌های احتمالی در نمونه نهایی ضروری خواهد بود. (رابینسون[2] و همکاران، ۲۰۱۷).

علاوه بر تحلیل و گزارش اطلاعات جمعیت‌شناختی که می‌تواند داخل متن یا با استفاده از جدول انجام شود لازم است اطلاعات توصیفی مربوط به سؤال‌ها گزارش شوند. در روانسنجی، شاخص‌های توصیفی سؤال‌ها نقش مهمی در ارزیابی کیفیت و عملکرد هر سؤال دارند. این شاخص‌ها به پژوهشگر کمک می‌کنند تا سؤال‌های ضعیف، مبهم یا نامناسب را شناسایی کرده و ابزار را بهینه‌سازی کند. مهم‌ترین شاخص‌های توصیفی پیشنهاد شده در این پژوهش ترکیبی از توصیه‌های دویلز (۲۰۱۷)، استرینر و همکاران (۲۰۱۴) و فور (۲۰۲۱) برای اطلاعات مورد نیاز در گزارش جدول توصیفی پیشنهاد شده است. میانگین که نشان‌دهنده سطح پاسخ‌دهی کلی به هر سؤال است. سؤال‌هایی با میانگین بسیار بالا یا پایین ممکن است سوگیری داشته باشند و باید محتوای آنها بررسی شود. این بررسی در روایی محتوایی، میزان اشتراک و جایگاه سؤال در ساختار عاملی و همبستگی سؤال با نمره کل می‌تواند صورت گیرد. انحراف معیار میزان پراکندگی پاسخ‌ها را نشان می‌دهد. انحراف معیار پایین ممکن است نشان‌دهنده عدم تنوع پاسخ‌ها باشد. در این صورت افراد نسبتا مشابه پاسخ داده‌اند و سؤال امکان تولید پاسخ‌های متنوع را نداشته است. کجی[3] نشان می‌دهد که توزیع پاسخ‌ها به سمت کدام طرف متمایل است. میزان کجی بین ۱- تا ۱+ مطلوب است. با وجود این بین ۲- تا ۲+ قابل قبول ارزیابی می‌شوند. در صورتی که مقدار کجی خارج از بازه قابل قبول باشد رفتار سؤال باید در ارزیابی‌های مختلف بررسی شود. کشیدگی[4] میزان تمرکز پاسخ‌ها در اطراف میانگین را نشان می‌دهد. تفسیر کشیدگی هم مثل کجی است اما اهمیت کمتری نسبت به آن دارد. شاخص تمایز[5] توانایی سؤال در تفکیک افراد با نمرات بالا و پایین را نشان می‌دهد. این شاخص در ابزارهایی که پاسخ درست تعریف می‌شود مهم‌تر است. در نظریه کلاسیک آزمون برای ابزارهای نگرش‌سنج و شخصیت که پاسخ درست تعریف نمی‌شود همبستگی سؤال با نمره کل به عنوان شاخص تمییز هم تفسیر می‌شود. حدقل قابل قبول برای این شاخص توسط کلاین (۲۰۱۳) و نانالی و برنشتاین[6] (۱۹۹۴) ۰/۳۰ توصیه شده است. در این پژوهش با توجه به وابستگی ضریب همبستگی به حجم نمونه و پراکندگی دو معیار معنی‌داری و حداقل ۰/۳۰ معیار رابطه سؤال با نمره کل توصیه می‌شود. درصد پاسخ‌دهی به گزینه‌ها بررسی می‌کند که آیا به‌طور متوازن انتخاب شده‌اند یا خیر. گزینه‌هایی که به‌ندرت انتخاب می‌شوند ممکن است نیاز به بازنگری داشته باشند. زمانی که مقیاس پاسخ مثل لیکرت برای همه سؤال‌ها ثابت است اگر گزینه‌ای درصد انتخاب پایینی

[1] Maslovskaya
[2] Robinson
[3] Skewness
[4] Kurtosis
[5] Item discrimination
[6] Nunnally & Bernstein

داشته باشد باید محتوای سؤال بازبینی شود. نکته مهم اینکه در روانسنجی بین‌فرهنگی هم ممکن است سؤالی حذف شود. اما باید توجه کرد که این آخرین راهکار است و زمانی انجام می‌شود که سؤال در چندین ارزیابی مختلف از جمله روایی محتوایی و تحلیل عاملی نامطلوب ارزیابی شود. در گزارش پژوهش هم باید بحث مفصل اقناعی درباره حذف سؤال ارائه شود.

ارزیابی اعتبار

اعتبار ابزار اندازه‌گیری دارای دو وجه دقت و پایایی است. کو و لی[1] (۲۰۱۶) چهار نوع اصلی اعتبار را از هم تفکیک کرده است: پایایی بازآزمایی[2]، توافق بین‌داوران[3]، پایایی فرم‌های موازی[4]، و همگونی درونی[5]. پایایی بازآزمایی میزان ثبات یک آزمون را در صورت اجرای مجدد بر روی همان نمونه در زمان‌های مختلف اندازه‌گیری می‌کند و برای موقعیت‌هایی که ثبات نتایج ابزار در طول زمان اهمیت دارد مناسب است. توافق بین داوران، میزان سازگاری و توافق میان دو یا چند ارزیاب یا ناظر در ارزیابی‌ها، قضاوت‌ها یا امتیازدهی‌هایشان نسبت به یک پدیده یا رفتار خاص را بررسی می‌کند، پایایی فرم‌های موازی به همبستگی بین دو ابزار که برای سنجش یک متغیر طراحی شده‌اند اشاره دارد، و پایایی درونی میزان همبستگی بین سؤال‌های یک آزمون را که هدف‌شان سنجش یک سازه واحد است، بررسی می‌کند (نانالی و برنشتاین، ۱۹۹۴).

همچنین در مواردی هم دو نوع اعتبار درونی و بیرونی معرفی شده است (بردهوشی و ارفورد[6]، ۲۰۱۷). اعتبار درونی به این معنی است که ابزار در درون خود ثبات دارد. به عبارت دیگر، اگر یک سؤال به شکل متفاوتی مطرح شود، نتایج یکسانی به دست می‌دهد. آلفای کرونباخ رایج‌ترین روش ارزیابی این اعتبار است. اعتبار بیرونی به چگونگی مقایسه نتایج با نتایج بین افراد یا در طول زمان اشاره دارد. این نوع اعتبار اغلب با استفاده از روش‌های بازآزمایی، توافق بین ارزیابان و فرم‌های موازی اندازه‌گیری می‌شود. پاشاشریفی (۱۴۰۰) این دو رویکرد اعتبار را به دو نوع منبع خطای ابزار نسبت می‌دهد.

آلفای کرونباخ: ضریب آلفای کرونباخ[7] (α) سازگاری درونی ابزار یا قدرت روابط بین سؤال‌ها را نشان می‌دهد. آلفا کرونباخ، کوواریانس بین سؤال‌ها یا چگونگی تغییر سؤال A هنگام تغییر سؤال B را توصیف می‌کند (کرونباخ، ۱۹۵۱). نمرات معمولاً از ۰ تا ۱ متغیر هستند، اما مقدار α منفی می‌تواند زمانی رخ دهد که سؤال‌ها همبستگی مثبت نداشته باشند (اورساچی[8] و همکاران، ۲۰۱۵). مقدار آلفای ۰/۷ تا ۰/۸ اغلب برای علوم رفتاری مطلوب در نظر گرفته می‌شود، اما این در ابزارهایی که حاوی سؤال‌های زیادی هستند می‌تواند به طور تصنعی افزایش یابد. کومار[9] (۲۰۲۴) ادعا می‌کند کرونباخ هرگز معیار ۰/۷ را مطرح نکرده است و احتمالا این نقطه برش برای اولین بار توسط نانالی و برنشتاین (۱۹۹۴) پیشنهاد شده است. برخی منابع مقدار آلفای بالای ۰/۹۰ را مطلوب‌تر توصیف کرده‌اند (مک‌نیش[10]، ۲۰۱۸) اما برخی آن را نشانه ارتباط زیاد چند سؤال می‌دانند؛ در این حالت چند سؤال یک چیز را می‌سنجند و افزوده اختصاصی ندارند (سوان و همکاران، ۲۰۲۳). با این حال، تابر[11] (۲۰۱۸) آستانه قابل قبول برای مقدار آلفا را به هدف و حساسیت نتایج اندازه‌گیری مربوط می‌داند. به عنوان مثال، شواهد اعتبار باید برای آزمون‌های پرخطر و آزمون‌های شناختی (مثلاً آزمون‌های هوش یا آزمون‌های ورود به دانشگاه) قوی‌تر از ابزارهای غربالگری نگرشی و شخصیتی (مثلاً پرسشنامه‌های علایق یا آزمون‌های شخصیت غیرتشخیصی) باشد. همچنین وقتی مقدار آلفا خیلی پایین باشد، نشان می‌دهد که سؤال(ها) نامناسب یا به هم نامربوط هستند. هوسی[12] و همکاران (۲۰۲۳) نتیجه گرفته‌اند که فراوانی مقادیر بالای ۰/۷ به شکل مشکوکی زیاد است. این پژوهشگران توصیه می‌کنند حداقل مقدار آلفا مشابه با سطح معنی داری الزام تعریف نشود تا بهره‌برداران با مقادیر واقعی‌تری مواجه شوند. زمانی که ابزار اندازه‌گیری چندعاملی است گزارش آلفای کرونباخ برای کل مقیاس غلط است. اما اگر نمره عامل‌ها برای محاسبه نمره کل منطقاً جمع پذیر باشند آلفا هم برای خرده‌مقیاس‌ها و هم برای نمره کل گزارش می‌شود.

با وجود محبوبیت بالای آلفای کرونباخ اما انتقادهایی هم به آن وارد شده است. برای مثال کومار (۲۰۲۴) بیان می‌کند که کرونباخ هرگز آلفا را به عنوان یک شاخص اعتبار پیشنهاد نکرده است بلکه آن را به عنوان یک معیار جایگزین برای تعادل‌یابی پایایی بازآزمایی پیشنهاد کرده است. دان[13] و همکاران (۲۰۱۴) به تأثیر رعایت نشدن مفروضه‌های این روش به ویژه فرض معادل‌بودن تاو[14] در کم‌برآورد ضریب آلفا اشاره کرده است. مفروضه معادل بودن تاو به معنای برابری مشارکت همه سؤال‌ها در برآورد نمره کل است. طبق این سیجسما[15] (۲۰۰۸) آلفای کرونباخ را به عنوان شاخص مناسب اعتبار رد می‌کند و آن را حد پایین اعتبار قلمداد می‌کند.

روش‌های مختلفی برای جایگزینی آلفای کرونباخ معرفی شده است که از این بین اومگای مک‌دونالد[16] (مک‌دونالد، ۱۹۹۹) در سال‌های اخیر بسیار پیشنهاد شده و به کار رفته است. این شاخص از شاخص‌های اعتبار مرکب[17] یا اعتبار سازه[18] است که از چارچوب تحلیل عاملی برای برآورد اعتبار استفاده می‌کند (کومار، ۲۰۲۴). امگای مک‌دونالد زمانی که معادل‌بودن تاو برآورده نشود، مفید است زیرا اجازه می‌دهد ارتباط بین هر سؤال و کل مقیاس تغییر

[1] Koo & Li
[2] Test-Retest
[3] Inter-Rater
[4] Alternative Forms
[5] Internal Consistency
[6] Bardhoshi & Erford
[7] Cronbach's Alpha Coefficients
[8] Ursachi
[9] Kumar
[10] McNeish
[11] Taber
[12] Hussey
[13] Dunn
[14] tau-equivalence
[15] Sijtsma
[16] McDonald's Omega
[17] Composite Validity
[18] Construct Validity

کند. دان و همکاران (۲۰۱۴) و ناجرا کاتالان[1] (۲۰۱۸) مجموعه‌ای از توصیه‌ها را برای تفسیر ω ارائه کرده‌اند که به طور کلی مقادیر بالای ۰/۷ مطلوب ارزیابی می‌شوند. با این حال، گزینه‌های دیگری از جمله فاصله اطمینان برای آلفا[2]، بزرگترین حد پایین[3] و ضریب H پیشنهاد شده‌اند. تشریح این روش‌ها فراتر از محدوده این مقاله است. برای مطالعه درباره آنها به مک‌نیش (۲۰۱۸)، بندرماخر[4] (۲۰۱۷) و کومار (۲۰۲۴) مراجعه شود.

پایایی بازآزمایی: در این مقاله برای Reliability معادل اعتبار استفاده شد؛ در حالی که برخی از منابع فارسی پایایی را پیشنهاد کرده‌اند. استدلال نویسنده مقاله حاضر این است که پایایی (معادل Stability) به ثبات و تکرارپذیری[5] نتایج به دست آمده از ابزار مربوط است و همان‌طور که تشریح شد این فقط یکی از وجوه Reliability است و نمی‌تواند شامل وجه دیگر آن یعنی دقت و همگونی[6] سؤال‌های ابزار شود. پایایی به این معناست که اگر یک ابزار اندازه‌گیری در دو یا چند زمان متفاوت در یک فرد اجرا شود، و در این فاصله وضعیت فرد در صفت مورد نظر تغییر نکرده باشد، باید نتایج مشابهی حاصل شود (آلدریج[7] و همکاران، ۲۰۱۷). اگر فاصله‌ی زمانی بین دو نوبت اجرای آزمون خیلی طولانی باشد، احتمال اینکه ویژگی مورد اندازه‌گیری واقعاً تغییر کرده باشد یا عوامل بیرونی مثل تغییرات خلقی، اتفاقات زندگی یا حتی خستگی و انگیزه اثر بگذارند می‌رود. برای همین، معمولاً فاصله زمانی دو روز تا دو هفته بین دو نوبت اجرا توصیه می‌شود. این بازه به اندازه‌ای کوتاه هست که فرد هنوز در همان وضعیت باقی مانده باشد، و به اندازه‌ای طولانی هست که اثر یادآوری سؤال‌ها به حداقل برسد (مارکس[8] و همکاران، ۲۰۰۳). برای محاسبه همبستگی بین دو اجرا ضریب همبستگی درون‌طبقه‌ای[9] مناسب‌تر از همبستگی پیرسون است. ضریب پیرسون میزان و جهت رابطه‌ی خطی بین دو متغیر را اندازه‌گیری می‌کند، در حالی که ضریب درون‌طبقه‌ای برای ارزیابی پایایی یا توافق میان اندازه‌گیری‌های تکرار شده در یک گروه به‌کار می‌رود (استرینر و کاتنر، ۲۰۱۴). طبق برخی منابع، مقدار همبستگی درون‌طبقه‌ای کمتر از ۰/۵ نشانه‌ی پایایی ضعیف، بین ۰/۵ تا ۰/۷۵ نشانه‌ی پایایی متوسط، بین ۰/۷۵ تا ۰/۹۰ پایایی خوب و بالاتر از ۰/۹۰ نشان‌دهنده‌ی پایایی عالی در نظر گرفته می‌شود (کو و لی، ۲۰۱۶ و لیجکست[10] و همکاران، ۲۰۱۹).

ارزیابی روایی

روایی یک ابزار اندازه‌گیری به توانایی آن در اندازه‌گیری«آنچه باید اندازه بگیرد» مربوط می‌شود (پاشاشریفی، ۱۴۰۰؛ سیف، ۱۴۰۲ و فور، ۲۰۲۱). بر خلاف اعتبار که وجوه مختلفی دارد، روایی یک مفهوم یگانه است که منابع مختلفی برای ارائه شواهدی برای آن وجود دارند (کالکبرنر، ۲۰۲۱)؛ از جمله روایی صوری، روایی محتوا، اعتبار مبتنی بر ملاک، و اعتبار سازه‌ای (فور، ۲۰۲۱).

روایی صوری: روایی صوری[11] به عنوان توانایی ابزار در قابل‌فهم بودن و مرتبط بودن برای جمعیت هدف تعریف می‌شود (یوسوف[12]، ۲۰۱۹). نوو[13] (۱۹۸۵) روایی صوری را به عنوان احساسات، نگرش‌ها و نظرات آزمون‌دهندگان، ارزیابان یا شرکت‌کنندگان نسبت به یک آزمون تعریف می‌کند.

در تعیین روایی صوری، روش کیفی مبتنی بر قضاوت متخصصان و شرکت‌کنندگان رایج تر است. با این حال، ارزیابی کمی روایی صوری نیز ممکن است. در روش کیفی تعیین روایی صوری با قضاوت متخصصان محتوایی و روانسنجی و البته شرکت‌کنندگان انجام می‌شود. در این ارزیابی مواردی مثل وضوح، ارتباط با هدف، منطقی بودن سؤال‌ها و استفاده از زبان مشترک (پرهیز از به کارگیری واژه‌های فنی و تخصصی) مورد بررسی می‌گیرند. پژوهش‌گر زمانی از روایی صوری ابزار اطمینان می‌یابد که اصلاحات مورد نظر گروه متخصصان و شرکت‌کنندگان اعمال و از نظر آنها تأیید شده باشد.

روایی صوری به صورت کمی می‌تواند با شاخص تأثیر سؤال اندازه‌گیری شود (یوسوف، ۲۰۱۹). طوری که برای هریک از سؤال‌های ابزار، طیف لیکرتی۵ نقطه‌ای طراحی می‌شود: کاملاً مهم است (امتیاز ۵)، مهم است (امتیاز ۴)، به طور متوسط مهم است (امتیاز ۳)، اندکی مهم است (امتیاز ۲) و اصلاً مهم نیست (امتیاز۱). سپس ابزار جهت تعیین روایی صوری در اختیار نمونه‌ای از افراد گروه هدف قرار می‌گیرد و پس از تکمیل با فرمول تأثیر سؤال[14] (فرمول ۱)، روایی صوری برای هر سؤال محاسبه می‌شود.

فرمول۱: نمره تأثیر $$\text{Impact Score} = \frac{\mathbf{n_e \times \sum Fx}}{\mathbf{N^2}}$$

در این فرمول:

n_e تعداد ارزیاب‌هایی است که به سؤال نمره ۴ یا ۵ داده‌اند،

F تعداد افرادی که گزینه مورد نظر را انتخاب کرده‌اند،

X امتیاز گزینه مورد نظر که مثلا امتیاز «گزینه اصلاً مهم نیست» برابر با ۱ است.

[1] Nájera Catalán
[2] alpha with confidence interval
[3] greatest lower bound
[4] Bendermacher
[5] Replicability
[6] Consistency
[7] Aldridge
[8] Marx
[9] Intraclass Correlation Coefficient
[10] Liljequist
[11] Face Validity
[12] Yusoff
[13] Nevo
[14] Impact Score furmula

برای مثال اگر ۷ نفر گزینه «کاملا مهم است»، ۶ نفر گزینه «مهم است»، ۲ نفر گزینه «متوسط مهم است»، ۴ نفر گزینه «اندکی مهم است» و ۱ نفر گزینه «اصلاً مهم نیست» را انتخاب کنند، n_e می‌شود ۱۳؛ $\sum FX$ برابر با ۰/۶۵ و نمره تأثیر برابر با ۲/۴۱ به دست می‌آید. حداقل مقدار قابل قبول برای نمره تأثیر ۱/۵ پیشنهاد شده است (هاردسی و بردن[1]، ۲۰۰۴).

روایی محتوایی: روایی محتوی به طور معمول به این سؤال پاسخ می‌دهد که ابزار طراحی شده تا چه حد جوانب مهم و اصلی مفهوم مورد اندازه‌گیری را در بردارد. این ارزیابی نیز مشابه روایی صوری به دو صورت کیفی و کمی قابل ارزیابی است. در ارزیابی کیفی روایی محتوایی مواردی مثل رعایت دستور زبان، استفاده از کلمات مناسب، اهمیت سؤال‌ها، قرارگیری سؤال‌ها در جای مناسب خود، زمان تکمیل ابزار طراحی شده از طرف گروهی از متخصصان موضوعی مورد توجه قرار می‌گیرد. روایی محتوی کمی نیز بر اساس نظرات متخصصین و با محاسبه دو شاخص نسبت روایی محتوایی[2] (CVR) و شاخص روایی محتوایی[3] (CVI) محاسبه می‌شود. نسبت روایی محتوایی ضرورت سؤال و شاخص روایی محتوا کیفیت طراحی سؤال را ارزیابی می‌کند. برای محاسبه این شاخص‌ها ارزیاب‌ها به هر سؤال ابزار، در خصوص ویژگی‌هایی مثل ضروری یا مفید بودن (CVI)یا اختصاصی بودن، سادگی و روان بودن و یا شفاف بودن سؤال (CVR)، در طیف لیکرت ۴ نقطه‌ای نظر می‌دهند. مثلاً جهت معیار ضروری بودن گزینه‌های ضروری نیست =۱، نسبتاً ضروری است = ۲، و ضروری است = ۳ و کاملاً ضروری است = ۴، طرح می‌شوند.

فرمول ۲: شاخص روایی محتوایی $$CVI = \frac{ne_{3,4}}{\frac{N}{2}}$$

در فرمول ۲ $ne_{3,4}$ تعداد ارزیاب‌هایی که به سؤال نمره ۳ یا ۴ داده‌اند و N تعداد کل ارزیاب‌هاست.

فرمول ۳: نسبت روایی محتوایی $$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

در فرمول ۳، ne تعداد ارزیاب‌هایی است که گویه مورد نظر را ضروری یا سودمند می‌دانند و N تعداد کل ارزیابان یا داورانی است که یک گویه را بررسی کرده‌اند.

در تفسیر CVI مقادیر ۰/۷۹ یا بیشتر نشان‌دهنده مطلوب بودن سؤال است. مقادیر ۰/۷۰ تا ۰/۷۹ نیاز به بازبینی دارند و کمتر از ۰/۷۰ نشان دهنده مشکل اساسی در سؤال است. اگر تعداد اعضای خبرگان بیش از ۵ نفر باشند حداقل قابل قبول ۰/۷۸ منظور می‌شود (لین[4]، ۱۹۸۶ و پایلوت و بک[5]، ۲۰۰۶). همچنین پایلوت و بک (۲۰۰۶) دو روش را برای محاسبه روایی محتوایی کل ابزار بر اساس CVI پیشنهاد کرده‌اند. اول محاسبه میانگین CVI برای همه سؤال‌ها و دوم محاسبه نسبت سؤال‌هایی که توسط داوران ۳ یا ۴ گرفته‌اند. تفسیر CVR بر اساس جدول لاوشه[6] (آیری و اسکالی[7]، ۲۰۱۴) صورت می‌گیرد. در این جدول حد قابل قبول بر اساس تعداد اعضای خبرگان تعیین می‌شود. جدول ۱ معیار لاوشه را برای تعداد اعضای مختلف خبرگان نشان می‌دهد.

روایی ملاکی: روایی ملاکی[8] میزان تطابق نتیجه ابزار اندازگیری مورد تحلیل با یک معیار غیرآزمونی[9] از پیش‌تعریف‌شده که استاندارد طلایی[10] خوانده می‌شود را ارزیابی می‌کند (فور، ۲۰۲۱). نکته‌ای که در این تعریف توسط صاحب‌نظران تاکید شده است غیرآزمونی بودن ملاک است (کین[11]، ۱۹۹۲ و کالکبرنر، ۲۰۲۱). اما همچنان در پژوهش‌های زیادی همبستگی نمره ابزار با یک ابزار مشابه دیگر به عنوان روایی ملاک گزارش می‌شود که در بخش بعدی توضیح داده خواهد شد که این همبستگی روایی همگرا است. این روایی در واقع نشان می‌دهد که ابزار تا چه حد نتایج واقعی‌ای را که برای سنجش آن طراحی شده، به‌درستی بازتاب می‌دهد. در روایی همزمان[12] ارتباط نمره‌های ابزار با یک ملاک غیرآزمونی در زمان حال بررسی می‌شود. به‌عنوان مثال، اگر یک روان‌سنج نمره‌های افراد در یک پرسشنامه‌ی افسردگی را با نمره‌های حاصل از مصاحبه بالینی روان‌پزشک را مقایسه کند، در حال بررسی روایی ملاکی همزمان ابزار است. اگر همبستگی دو مجموعه نمره بالا باشد، شواهدی از روایی همزمان ابزار فراهم می‌کند، زیرا نمره‌ها با ملاک غیرآزمونی (ارزیابی روانشناس) هم‌راستا هستند. شواهد روایی ملاکی همچنین می‌توانند شامل روایی پیش‌بین[13] باشد، یعنی میزانی که نمره‌های حاصل از ابزار اندازه‌گیری بتوانند ملاکی غیرآزمونی در آینده یا حتی گذشته را پیش‌بینی کنند. برای نمونه، یک روان‌سنج ممکن است روایی پیش‌بین یک ابزار آمادگی شغلی را با بررسی میزان پیش‌بینی نمره‌های آمادگی در ارزیابی‌های عملکرد شغلی آینده توسط کارفرمایان، تحلیل کند (بندالوس و فینی، ۲۰۱۹). در صورتی که نمره ملاک پیوسته باشد ضریب همبستگی و زمانی که دوارزشی (مثلا بیمار/سالم یا قبول/رد) باشد محاسبه شاخص‌های حساسیت[14] و ویژگی[15] توصیه می‌شود (ساکیل[16] و همکاران، ۲۰۰۶ و سوان و همکاران، ۲۰۲۳). حساسیت نشان‌دهنده‌ی توانایی یک ابزار در شناسایی درست افراد دارای صفت

1 Hardesty & Bearden
2 Content Validity Ratio
3 Content Validity Index
4 Lynn
5 Polit & Beck
6 Lowshe
7 Ayre & Scally
8 Criterion Validity
9 non-test criterion
10 Golden Standard
11 Kane
12 Concurrent Validity
13 Predictive Validity
14 Sensitivity
15 Specificity
16 Sockeel

مورد نظر (مثلاً بیماری) است. به عبارت دیگر، نسبت کسانی که واقعاً بیمارند و ابزار آنها را به‌درستی به‌عنوان مثبت شناسایی می‌کند. ویژگی بیانگر توانایی آزمون در شناسایی درست افراد بدون صفت مورد نظر است (مثلاً افراد سالم) (موناگان[1] و همکاران، ۲۰۲۱). این دو شاخص (حساسیت و ویژگی) نسبت معکوس با یکدیگر دارند؛ طوری که با افزایش یکی، دیگری کاهش می‌یابد. درصد قابل قبول برای هر شاخص بسته به هدف مورد نظر از ابزار اندازه‌گیری متفاوت است. در ملاک‌های پرخطر ولی پیامدهای کم‌خطر، حساسیت بیشتر ممکن است مطلوب‌تر باشد. در مقابل، ویژگی بالاتر در ملاک‌های با خطر پایین یا پیامدهای با خطر بالا ترجیح داده می‌شود (موناگان، ۲۰۲۱).

جدول ۱

معیار لاوشه حداقل مقادیر قابل قبول برای CVR

حداقل مقدار نسبت روایی محتوا	تعداد اعضای خبرگان
۰/۹۹	۵
۰/۹۹	۶
۰/۹۹	۷
۰/۷۵	۸
۰/۷۸	۹
۰/۶۲	۱۰
۰/۴۹	۱۵
۰/۴۲	۲۰
۰/۳۷	۲۵
۰/۳۳	۳۰
۰/۳۱	۳۵
۰/۲۹	۴۰

روایی سازه

روایی سازه به درجه‌ای اطلاق می‌شود که در آن یک ابزار اندازه‌گیری، توانایی سنجش دقیق و معتبر سازه‌ای نظری یا فرضی را دارد که هدف طراحی آن بوده است (فور، ۲۰۲۱). این نوع روایی، نشان‌دهنده میزان هم‌خوانی بین نتایج حاصل از ابزار و ویژگی‌های نظری سازه مورد نظر است. در واقع، روایی سازه تأیید می‌کند که آیا ابزار مورد استفاده، واقعاً ماهیت مفهوم و انتزاعی همچون هوش، اضطراب، یا ویژگی‌های شخصیتی را اندازه می‌گیرد یا صرفاً بازتاب‌دهنده متغیرهای ناخواسته و بیرونی است (ریکوف و مارکولیدس[2]، ۲۰۱۱). روایی سازه به اندازه‌ای جامع است که در تعریف معادل مفهوم کلی روایی است. در این پژوهش روایی سازه در سه دسته معرفی شده است: رابطه با ابزارهای معتبر قبلی، تحلیل عاملی و مقایسه در گروه‌های شناخته شده.

رابطه با ابزارهای معتبر قبلی: در این روش روایی همگرا[3] و تشخیصی[4] ابزار ارزیابی می‌شوند. روایی همگرا زمانی برقرار است که بین اندازه‌گیری‌های مختلف از یک سازه واحد رابطه وجود داشته باشد. برای مثال، مقیاس جدید افسردگی زمانی از روایی همگرا برخوردار است که نمره حاصل از آن با نمره حاصل از یک ابزار شناخته‌شده برای غربالگری افسردگی، مانند پرسشنامه‌ی بک، رابطه داشته باشد. در برخی منابع حداقل مقدار ۰/۵ برای این همبستگی پیشنهاد شده است (سوانک و مولن[5]، ۲۰۱۷ و چئونگ[6] و همکاران، ۲۰۲۴). طبق تعریف روایی تشخیصی یا واگر[7] ابزار معرفی شده باید با ابزارهای سازه‌های نظری نامرتبط در همان حوزه ارتباط کم یا کمتری داشته باشد. با توجه به مثال پیشین، برای ارزیابی روایی تشخیصی ابزار جدید افسردگی باید رابطه نمره‌های حاصل از آن با نمره‌های یک ابزار معتبر سنجش اضطراب مانند پرسش‌نامه‌ی اضطراب بک بررسی شود. طبق نظریه شناختی بک در روان‌شناسی انتظار می‌رود همبستگی بین افسردگی و اضطراب در حد ضعیف تا متوسط باشد (وانگ و گورستین[8]، ۲۰۱۳). این میزان همبستگی طبق توصیه سوانک و مولن (۲۰۱۷) باید کمتر از ۰/۴ باشد. به نظر می‌رسد توجه صرف به این مقادیر همبستگی می‌تواند گمراه‌کننده باشد. چرا که ارزیابی ضرایب همبستگی بدون توجه به شاخص‌هایی مثل حجم نمونه و پراکندگی متغیرها دچار کاستی است. چنانچه ممکن است در نمونه کوچک همبستگی ۰/۶ معنی‌دار نباشد و در نمونه بزرگ همبستگی ۰/۱۰ معنی‌دار باشد. برای همین استفاده از شاخص‌هایی مثل سطح‌معنی‌داری یا اندازه اثر مناسب‌تر هستند. نویسنده برای روایی همگرا دو معیار ضریب همبستگی بالاتر از ۰/۵۰ و همزمان معنی‌داری آماری را پیشنهاد می‌کند. درباره روایی تشخیصی نکته مهم مراجعه به نظریه است. در مثال بالا نظریه از رابطه کم و متوسط افسردگی و اضطراب حمایت می‌کرد؛ طبق همان نظریه رابطه افسردگی و وسواس باید کمتر از متوسط باشد. بنابراین در نظر گرفتن پیش‌بینی‌های نظریه مهمترین راهبرد در تعیین سطح رابطه مورد انتظار برای روایی تشخیصی است.

[1] Monaghan
[2] Raykov & Marcoulides
[3] Convergent Validity
[4] Discriminant Validity
[5] Swank & Mullen
[6] Cheung
[7] Divergent Validity
[8] Wang & Gorenstein

تحلیل عاملی: تحلیل عاملی یک روش آماری در روانسنجی است که بر اساس روابط بین سؤال‌ها ابعاد زیربنایی ابزار اندازه‌گیری را توضیح می‌دهد (کلاین، ۲۰۱۴). این روش به محققان کمک می‌کند تا با کاهش تعداد متغیرها و گروه‌بندی سؤال‌هایی که سازه یکسانی را اندازه‌گیری می‌کنند، ساختار یک آزمون را درک کنند (شرستا، ۲۰۲۱). این فرآیند با اطمینان از اینکه سؤال‌ها به طور مرتبطی منعکس‌کننده سازه‌های مورد نظر هستند، روایی ارزیابی‌ها را تضمین می‌کنند. تحلیل عاملی با دو روش اکتشافی[1] و تأییدی[2] انجام می‌شود. تحلیل عاملی اکتشافی برای زمانی مناسب است که روان‌سنج شناخت و فرضیه‌ای برای رابطه سؤال با سازه زیربنایی نداشته باشد و با این روش برای رسیدن به یک ساختار عاملی که در آن رابطه سؤال‌ها با عامل‌ها مشخص می‌شود تلاش می‌کند. اما در تحلیل عاملی تأییدی یک ساختار عاملی مفروضی وجود دارد که روان‌سنج قصد دارد در یک نمونه جدید برازش آن را بررسی کند (کلاین، ۲۰۱۴). تحلیل عاملی اکتشافی در ساخت ابزارهای اندازه‌گیری و تحلیل عاملی تأییدی در ارزیابی ابزارهای قبلا ساخته شده در نمونه جدید مناسب هستند. بنابراین تحلیل عاملی تأییدی برای روانسنجی بین فرهنگی مناسب‌تر است.

یکی از موضوعات مهم در تحلیل عاملی تأییدی انتخاب روش برآورد است. رایج‌ترین روش، بیشینه درست‌نمایی[3] است. استفاده از این روش مفروضه‌هایی مثل نرمال بودن چندمتغیری، خطی بودن رابطه نشانگرها با سازه‌ها، پیوسته بودن متغیرها و استقلال مشاهده‌ها دارد. نرمال بودن چند متغیری با آزمون‌های مختلفی از جمله مردیا[4] (۱۹۸۰) بررسی می‌شود. در صورت تأیید فرض صفر کجی و کشیدگی انحراف معنی‌داری با توزیع نرمال ندارند. برآورد کمترین مربعات وزنی با تعدیل میانگین و واریانس[5] (WLSMV) یکی از روش‌های جایگزین برای شرایطی است که مفروضه نرمال‌بودن چندمتغیری رد شود. همچنین این روش برای زمانی که نمره متغیرها به جای پیوسته طبقه‌ای مثل لیکرت هستند یا حجم نمونه مکفی نیست مناسب‌تر است (کلاین، ۲۰۲۳).

دو نوع ارزیابی در تحلیل عاملی تأییدی انجام می‌شود. اول رابطه سؤال با عامل زیربنایی و دوم شاخص‌های برازش[6] است. درباره حداقل مقدار بار عاملی پیشنهادهای مختلفی ارائه شده است که در این پژوهش باز هم معنی‌داری آماری به جای آنها پیشنهاد می‌شود. شاخص‌های برازش ارزیابی می‌کنند که مدل اولیه تا چه اندازه با داده‌های مشاهده‌شده مطابقت دارد. ارزیابی برازش مدل بخش مرکزی در تحلیل عاملی تأییدی و اغلب در بررسی روایی ابزارهای روان‌شناختی به شمار می‌آید (گورتزکو[7] و همکاران، ۲۰۲۴). رایج‌ترین شاخص‌های برازش که در تحلیل‌عاملی تأییدی و همچنین مدل‌معادلات‌ساختاری عبارت‌اند: خی‌دو، نسبت خی‌دو به درجه آزادی، شاخص برازش تطبیقی[8] (CFI)، شاخص تاکر–لوئیس[9] (TLI / NNFI)، میانگین مربع خطای تقریب[10] (RMSEA)، ریشه میانگین مربع باقی‌مانده استانداردشده[11] (SRMR)، شاخص برازش عمومی[12] (GFI)، شاخص برازش عمومی تعدیل‌شده[13] (AGFI)، شاخص برازش نرمال‌شده[14] (NFI)، شاخص برازش افزایشی[15] (IFI)، شاخص‌های برازش صرفه‌جویانه[16] (PGFI / PNFI / PCFI). تصمیم‌گیری درباره این شاخص‌ها مبتنی بر معیاری مثل سطح معنی‌داری نیست و نقطه‌برش‌های مختلفی در منابع گزارش شده است که یکی از چالش‌های اساسی در تصمیم‌گیری روایی ابزارها است. در جدول ۲ نقطه برش برای رایج‌ترین شاخص‌های برازش طبق چهار منبع از پراستنادترین منابع خلاصه شده است. نکته اینکه خی‌دو به حجم نمونه حساس است و در نمونه‌های بزرگ اعتبار زیادی ندارد و بهتر است به شاخص‌های دیگر توجه شود. در سال‌های اخیر انتقاداتی به این نقطه برش‌های ثابت وارد شده است (گورتزکو و همکاران، ۲۰۲۴). از جمله مک‌نیش و ولف[17] (۲۰۲۳) معتقدند استفاده از نقاط برش ثابت ممکن است گمراه‌کننده باشد، زیرا این نقاط برش فقط برای یک نوع مدل خاص (مدل سه‌عاملی با بارهای خاص) استخراج شده‌اند و به مدل‌های دیگر قابل تعمیم نیستند. آنها روشی مبتنی بر شبیه‌سازی پیشنهاد می‌کنند که نقاط برش را متناسب با ویژگی‌های خاص مدل (تعداد عوامل، نوع داده، بارهای عاملی، تعداد سؤال‌ها و ...) تعیین می‌کند. این روش با استفاده از نرم‌افزار تحت وب و متن‌باز قابل اجراست و نیاز به دانش برنامه‌نویسی ندارد. به طور مشابه گروسکلت[18] و همکاران (۲۰۲۴) استدلال می‌کنند که نقطه‌برش‌های گزارش شده برآمده از مطالعات شبیه‌سازی شده هستند و هشدار داده‌اند که این مقادیر فقط در شرایطی معتبر هستند که مشابه با سناریوهای شبیه‌سازی اولیه باشند. آنچه مسلم است اینکه این شاخص‌ها انعطاف‌پذیر هستند.

روایی گروه‌های شناخته‌شده[19]: این روایی به میزان توانایی یک ابزار اندازه‌گیری در تشخیص تفاوت بین گروه‌هایی اشاره دارد که انتظار می‌رود در ویژگی مورد سنجش با هم تفاوت داشته باشند. به زبان ساده‌تر، اگر یک آزمون برای سنجش یک ویژگی خاص طراحی شده باشد، زمانی دارای روایی گروه‌های شناخته‌شده است که بتواند افرادی را که به‌طور قطعی آن ویژگی را دارند، از کسانی که آن را ندارند، به‌درستی متمایز کند (شولتز[20]، ۲۰۲۰ و

[1] Exploratory Factor Analysis
[2] Confirmatori Factor Analysis
[3] Maximum likelihood
[4] Mardia
[5] Weighted Least Squares Mean and Variance adjusted
[6] Goodness-of-Fit Indices
[7] Goretzko
[8] Comparative Fit Index
[9] Tucker-Lewis Index/Non-Normed Fit Index
[10] Root Mean Square Error of Approximation
[11] Standardized Root Mean Square Residual
[12] Goodness of Fit Index
[13] Adjusted Goodness of Fit Index
[14] Normed Fit Index
[15] Incremental Fit Index
[16] Parsimony Goodness / Normed / Comparative Fit Index
[17] McNeish & Wolf
[18] Groskurth
[19] known-groups validity
[20] Shultz

فور، ۲۰۲۱). برای مثال روانسنجی که برای ارزیابی روایی مقیاس اعتیاد به اینترنت انتظار دارد گروهی که ساعت‌های زیادی را در استفاده از اینترنت می‌گذارنند میانگین بیشتری از گروهی که زمان کمتری صرف این کار می‌کنند به دنبال روایی گروه‌های شناخته شده است. او این روایی را با معنی‌دار شدن آزمون آماری t برای مقایسه دو گروه مستقل انجام می‌دهد. دقت این روش وابسته به کیفیت تعریف گروه‌هاست (دیویدسون[1]، ۲۰۲۴).

جدول ۲

نقطه برش شاخص‌های برازش در منابع مختلف

شاخص	منبع			
	هو و بنتلر (۱۹۹۹)	مارش و همکاران (۲۰۰۴)	هیر و همکاران (۲۰۰۶)	کلاین (۲۰۱۶)
χ^2	-	-	P>۰/۰۵	P>۰/۰۵
χ^2/df	-	۲ تا ۳	> ۳	> ۳
CFI	>۰/۹۵	>۰/۹۵	>۰/۹۵	>۰/۹۵
TLI / NNFI	>۰/۹۵	>۰/۹۵	>۰/۹۵	>۰/۹۵
RMSEA	<۰/۰۶	<۰/۰۸	<۰/۰۷	<۰/۰۸
SRMR	<۰/۰۸	<۰/۰۸	<۰/۰۸	<۰/۰۸
GFI	>۰/۹۰	-	>۰/۹۵	>۰/۹۵
AGFI	>۰/۹۵	-	>۰/۹۵	>۰/۸۵
NFI	>۰/۹۵	>۰/۹۵	>۰/۹۵	>۰/۹۵
IFI	>۰/۹۵	>۰/۹۵	>۰/۹۵	>۰/۹۵
PGFI / PNFI / PCFI	-	>۰/۵۰	>۰/۵۰	>۰/۵۰

بحث و نتیجه‌گیری

در این مقاله تلاش شد چارچوبی برای انجام و گزارش پژوهش‌های روانسنجی بین‌فرهنگی ارائه شود و پیشنهادهای عملی برای چالش‌های تصمیم‌گیری مبتنی بر منابع پراستناد و منابع اصلی ارائه شود. روانسنجی بین‌فرهنگی در این پژوهش با شش مرحله کلی شامل شامل انتخاب ابزار بر اساس مبانی نظری، انطباق بین‌فرهنگی، نمونه‌گیری و اجرا، تحلیل سؤال، ارزیابی اعتبار و ارزیابی روایی معرفی شد. این مراحل می‌تواند پژوهش‌های روانسنجی را مطمئن‌تر و قابل اعتمادتر و گزارش آنها را یک‌دست‌تر کند. همچنین معیاری برای ارزیابی گزارش‌های روانسنجی ابزارهای اندازه‌گیری است.

این پژوهش شامل محدودیت‌هایی نیز هست. اول اینکه اشاره‌ای به نظریه اندازه‌گیری نشده است و عمده مطالب مبتنی بر نظریه کلاسیک آزمون[2] نوشته شده‌اند. نظریه سؤال-پاسخ[3] مباحث متفاوتی درباره تحلیل‌های روانسنجی ارائه می‌دهد که در این مقاله پوشش داده نشده‌اند. اجرای چنین پژوهش‌هایی مستلزم استفاده از حجم نمونه بزرگ است و شواهد حاصل از آن در ابتدا مختص همان نمونه خواهد بود تا زمانی که پژوهشگران بعدی قابلیت تعمیم‌پذیری اعتبار و پایایی آن را اثبات کنند (کالکبرنر، ۲۰۲۱). علاوه بر این، ملاحظات متعدد دیگری در روانسنجی توسعه یافته‌اند که برای پژوهشگران علوم رفتاری مفید است اما در مقاله حاضر به آنها پرداخته نشده است؛ از جمله مصاحبه‌های شناختی[4] (این روش به پژوهشگران کمک می‌کند تا بفهمند آیا شرکت‌کنندگان سؤال‌ها را همان‌طور که طراح ابزار مدنظر داشته، درک می‌کنند یا نه، پترسون[5] و همکاران، ۲۰۱۷)، آزمون‌های تغییرناپذیری اندازه‌گیری[6] (بررسی می‌کنند آیا یک ابزار اندازه‌گیری در گروه‌های مختلف جمعیتی -مثلاً زنان و مردان، یا فرهنگ‌های مختلف- به شکل یکسانی عمل می‌کند یا نه، دیمیتروف[7]، ۲۰۱۰)، تحلیل عاملی تأییدی مرتبه دوم[8] (کرده[9] و همکاران، ۲۰۱۵). استفاده از ابزارها در خارج از نمونه‌های هنجاری (به‌کارگیری ابزارها در جمعیت‌هایی که با گروهی که ابزار برای آن طراحی و هنجاریابی شده، تفاوت دارند، هیز و وود[10]، ۲۰۱۷)، آزمون‌های حساس با پیامدهای مهم[11] (آزمون‌هایی که نتایج آنها تأثیر قابل توجهی بر زندگی فرد یا تصمیمات سازمانی دارند، بونه[12] و همکاران، ۲۰۱۳). به پژوهشگران علاقمند به راهنمانویسی توصیه می‌شود با مرور منابع مهم و روز دنیا به معرفی این رویکردها بپردازند و یا از آنها در پژوهش‌های روانسنجی در عمل بهره‌مند شوند.

همچنین این مطالعه محدود به روانسنجی بین‌فرهنگی است و موضوعات ابزارسازی را توضیح نمی‌دهد. بنابراین بخش‌هایی از روانسنجی کلاسیک مربوط به ساخت ابزارها مانند تحلیل عاملی اکتشافی، انتخاب سؤال‌ها از بانک سؤال، مقیاس‌سازی[13]، تعیین استاندارد[14] و ... خارج از هدف و ساختار این مطالعه قرار گرفتند. نکته دیگر اینکه در معرفی روایی و اعتبار به شاخص‌های رایج و موضوعات چالشی بسنده شد و برخی از شقوق آنها مثل ماترس

1 Davidson
2 Classical Test Theory
3 Item Response Theory
4 Cognitive Interviews
5 Peterson
6 Measurement Invariance
7 Dimitrov
8 Higher-order CFA
9 Credé
10 Hays & Wood
11 High-Stakes Tests
12 Boone
13 Scaling
14 Standard Setting

چندصفت–چندروش[1] یا معرفی نشدند. به علاقمندان توصیه می‌شود برای مطالعه بیشتر به منابع معرفی شده در این مقاله مراجعه کنند. این مقاله علاوه بر استفاده پژوهشگران و بهره‌برداران از ابزارهای روانسنجی می‌تواند برای مقاصد آموزشی در درس‌های مرتبط با روانسنجی نیز استفاده شود.

سپاس‌گزاری

از همکاران، دانشجویان و دوستانی که مطالب مفیدی درباره روانسنجی بین فرهنگی پیشنهاد کردند تشکر می‌شود.

ملاحظات اخلاقی

این مطالعه مروری بود و شامل شرکت‌کنندگان انسانی نبود. با وجود این نویسنده متعهد به موارد اخلاقی از جمله امانت داری و ذکر منابع معتبر بوده است.

حامی مالی

هزینه‌های انجام این پژوهش و نگارش مقاله به‌طور کامل توسط نویسنده تأمین شده و تحقیق حاضر از هیچ‌گونه منابع حمایتی برخوردار نبوده است.

مشارکت نویسندگان

همه فعالیت‌ها شامل جستجو و ارزیابی منابع، ترجمه، تدوین و نگارش مقاله بر عهده مجید یوسفی افراشته بوده است.

تعارض منافع

نویسنده اعلام می‌کند که هیچ‌گونه تعارض منافعی در این پژوهش وجود ندارد.

منابع

مارنات، گ. و رایت، ج. (۱۴۰۰). *راهنمای سنجش روانی*. ترجمه حمیدنصیری، حسن پاشاشریفی، محمدرضا نیکخو، مهدی گنجی و نسترن شریفی. ویراست ششم، رشد.

دبیرخانه کارگروه وزارتی اخلاق در پژوهش وزارت بهداشت. (۱۴۰۱). *مجموعه قوانین، دستورالعملها و راهنماهای اخلاقی در پژوهشهای زیست پزشکی ایران*، جهاد دانشگاهی

پاشاشریفی، ح و شریفی، ن. (۱۴۰۰). *اصول روانسنجی و روان‌آزمایی*. انتشارات رشد.

سیف، ع. ا. (۱۴۰۲). *اندازگیری، سنجش و ارزشیابی آموزشی*. نشر دوران.

References

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. *European Psychologist*, *22*(4), 207.

Arafat, S. Y., Chowdhury, H. R., Qusar, M. M. A. S., & Hafez, M. A. (2016). Cross cultural adaptation and psychometric validation of research instruments: a methodological review. *Journal of Behavioral Health*, *5*(3), 129-136.

Arifin, W. N. (2025). A Web-Based Sample Size Calculator for Structural Equation Modelling. *Education in Medicine Journal*, *17*(1).

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and evaluation in counseling and development*, *47*(1), 79-86.

Bandalos, D.L., & Finney, S.J. (2019). Factor analysis: Exploratory and confirmatory. In G.R. Hancock, L. M. Stapleton, & R.O. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 98-122). Routledge.

Bardhoshi, G., & Erford, B. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development. 50*(4), 256–263.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191.

[1] Multi-trait multi-method

Beavers, A. A., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, *18*(5/6), 1-13.

Bendermacher, N. (2017). An unbiased estimator of the greatest lower bound. *Journal of Modern Applied Statistical Methods*, *16*(1), 674–688.

Boone, W., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. Cogent Education, 4(1).

Borsboom, D. (2022). Possible futures for network psychometrics. *Psychometrika*, *87*(1), 253-265.

Borsboom, D., and Wijsen, L. D. (2017). Psychology's atomic bomb. *Assess. Educ.* 24, 440–446.

Bossuyt P.M., Reitsma J.B., Bruns D.E., Gatsonis C.A., Glasziou P.P., Irwig L.M., Lijmer J.G., Moher D., Rennie D. & de Vet H.C.W. (2003) Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. British Medical Journal 326, 41–44.

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia pacific journal of management*, *41*(2), 745-783.

Cicchetti D.V. (1999) Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions. Journal of Clinical and Experimental Neuropsychology 21(4), 567–570.

Cicchetti D.V. (2001) The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. Journal of Clinical and Experimental Neuropsychology 23(5), 695–700

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., ... & Bossuyt, P. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open*, *6*(11), e012799.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Lawrence Erlbaum

Credé, M., & Harms, P. (2015). 25 years of higher- order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. Journal of Organizational Behavior, 36(6), 845–872. Be, D., Whisman,

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Cruchinho, P., López-Franco, M. D., Capelas, M. L., Almeida, S., Bennett, P. M., Miranda da Silva, M., ... & Gaspar, F. (2024). Translation, cross-cultural adaptation, and validation of measurement instruments: a practical guideline for novice researchers. *Journal of Multidisciplinary Healthcare*, 2701-2728.

Dai, Y., Ding, J., Daveson, B. A., Chen, Y., Connolly, A., & Johnson, C. E. (2024). Validating performance status and activities of daily living assessment tools for Chinese palliative care in a cancer setting: A cross-cultural psychometric study. *Asia-Pacific Journal of Oncology Nursing*, *11*(12), 100613.

Davidson, M. (2024). Known-groups validity. In *Encyclopedia of quality of life and well-being research* (pp. 3764-3764). Cham: Springer International Publishing.

DeVellis, R. F. (2017). Scale Development: Theory and Applications (4th ed.). Thousand Oaks, CA: Sage.

Dimitrov, D. (2012). Statistical methods for validation of assessment scale data in counseling and related fields. American Counseling Association.

Dunn, T., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *The British Journal of Psychology*, *105*(3), 399–412.

Eich, E. (2014). Business not as usual. *Psychol. Sci.* 25, 3–6.

Feldt L.S. & Ankenmann R.D. (1999) Determining sample size for a test of the equality of alpha coefficients when the number of part–tests is small. Psychological Methods 4(4), 366–377.

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576.

Flores-Kanter, P. E., & Mosquera, M. (2023). How do you behave as a psychometrician? Research conduct in the context of psychometric research. *The Spanish Journal of Psychology*, *26*, e13.

Furr, R. M. (2021). *Psychometrics: an introduction*. SAGE publications.

Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, *84*(1), 123-144.

Groskurth, K., Bluemke, M., & Lechner, C. M. (2024). Why we need to abandon fixed cutoffs for goodness-of-fit indices: An extensive simulation and possible solutions. *Behavior Research Methods*, *56*(4), 3891-3914.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R*. (2006). Multivariate Data Analysis (6th ed.). Upper Saddle River, NJ Pearson Prentice Hall.*

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of business research*, *57*(2), 98-107.

Hays, D., & Wood, C. (2017). Stepping outside the normed sample: Implications for validity. Measurement and Evaluation in Counseling and Development. 50(4), 282–288.

Hernández, A., Hidalgo Montesinos, M. D., Hambleton, R. K., & Gómez Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema, 2020, vol. 32, num. 3, p. 390-398*.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in nursing & health*, *31*(2), 180-191.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2023). An aberrant abundance of Cronbach's alpha values at. 70. *PsyArXiv*.

Irwing, P., & Hughes, D. J. (2018). Test Development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, V.I (pp. 4-47). Hoboken, NJ: Wiley.

Isvoranu, A. M., Epskamp, S., Waldorp, L., & Borsboom, D. (Eds.). (2022). *Network psychometrics with R: A guide for behavioral and social scientists*. Taylor & Francis.

Kalkbrenner, M. T. (2021). A Practical Guide to Instrument Development and Score Validation in the Social Sciences: The MEASURE Approach. *Practical Assessment, Research & Evaluation*, *26*, 1.

Kalkbrenner, M. T. (2021). A Practical Guide to Instrument Development and Score Validation in the Social Sciences: The MEASURE Approach. *Practical Assessment, Research & Evaluation*, *26*, 1.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kanth, D. B., Indumathy, J., Kadhiravan, S., Nagasubramaniyan, G., & Sri Lekha, P. P. (2024). Introduction to Cultural Adaptations and Validation of Measurement Instruments. In *Measuring Couples and Family Dynamics in India: Cultural Adaptations and Validations* (pp. 1-16). Singapore: Springer Nature Singapore.

Kennedy, I. (2022). Sample size determination in test-retest and Cronbach alpha reliability estimates. *British Journal of Contemporary Education*, *2*(1), 17-29.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York: The Guilford Press.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE Life Sciences Education*, *18*(1)

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., ... & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies*, *48*(6), 661-671.

Kumar, R. V. (2024). Cronbach's alpha: Genesis, issues and alternatives. *Management*, *1*, 17.

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, *9*(11), 2531-2560.

Lacko, D., Čeněk, J., Točík, J., Avsec, A., Đorđević, V., Genc, A., ... & Subotić, S. (2022). The necessity of testing measurement invariance in cross-cultural research: Potential bias in cross-cultural comparisons with individualism–collectivism self-report scales. *Cross-Cultural Research*, *56*(2-3), 228-267.

Lenz, A., Gómez Soler, I., Dell'Aquilla, J., & Uribe, P. (2017). Translation and cross-cultural adaptation of assessments for use in counseling research. Measurement and Evaluation in Counseling and Development. 50(4), 224–231.

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PloS one*, *14*(7), e0219854.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*, *35*(6), 382-386.

Mardia, K. V. (1980). 9 tests of unvariate and multivariate normality. *Handbook of statistics*, *1*, 279-320.

Marnat, G. G., & Wright, J. (2016). *Handbook of psychological assessment*. (6th ed.; H. Nasiri, H. Pasha-Sharifi, M. Nik-khoo, M. Ganji, & N. Sharifi, Trans.). Roshd Publishing. [Persion]

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: evaluation of alternative estimation strategies and indicator construction. *Psychological methods*, *9*(3), 275.

Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of clinical epidemiology*, *56*(8), 730-735.

Maslovskaya, O., Struminskaya, B., & Durrant, G. (2022). The future of online data collection in social surveys: challenges, developments and applications. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 768-772.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412– 433.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412– 433.

McNeish, D., & Wolf, M. G. (2024). Direct discrepancy dynamic fit index cutoffs for arbitrary covariance structure models. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 835-862.

Meyer G.J. (2003) Guidelines for reporting information in studies of diagnostic accuracy: the STARD initiative. Journal of Personality Assessment 81(3), 191–193

Michell, J. (2008). Is psychometrics pathological science?. *Measurement*, *6*(1-2), 7-24.

Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., & Dmochowski, R. R. (2021). Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina*, *57*(5), 503.

Nájera Catalán, H. (2018). Reliability, population classification and weighting in multidimensional poverty measurement: A Monte Carlo Study. *Social Indicators Research*, *142*(3), 887–910.

Neal, Z., Neal, J. W., & Piteo, A. (2020). Call me maybe: Using incentives and follow-ups to increase principals' survey response rates. *Journal of Research on Educational Effectiveness*, *13*(4), 784-793.

Nunnally, J.C. and Bernstein, I.R. (1994), Psychometric theory, Ed. ke-3, McGraw-Hill, New York.

Paredes, J., & Carré, D. (2024). Looking for a broader mindset in psychometrics: the case for more participatory measurement practices. *Frontiers in Psychology*, *15*, 1389640.

Pasha-Sharifi, H., & Sharifi, N. (2021). *Principles of psychometrics and psychological testing* (6th ed.). Roshd Publishing. [Persion]

Peterson, C., Peterson, N., & Powell, K. (2017). Cognitive interviewing for item development: Validity evidence based on content and response Processes. Measurement and Evaluation in Counseling and Development. 50(4), 217–223.

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.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Robinson, J. K., McMichael, A. J., & Hernandez, C. (2017). Transparent reporting of demographic characteristics of study participants. *JAMA dermatology*, *153*(3), 263-264.

Saif, A. (2023). Measurement, assessment, and educational evaluation, (7th ed.). Doran Publishing. [Persion]

Salzberger, T. (2013). Attempting measurement of psychological attributes. Front. Psychol. 4:75.

Secretariat of the Ministerial Committee on Research Ethics, Ministry of Health. (2022). *Collection of laws, guidelines, and ethical instructions in biomedical research in Iran.* Jahad-e-Daneshgahi Publishing. [Persion]

Semkiv, I., Turetska, K., Kryvenko, I., & Kechur, R. (2022). Linguistic and psychometric validation of the Ukrainian translation of the Inventory of Personality Organization-Revised (IPO-R-UKR). *East European Journal of Psycholinguistics*, *9*(1), 176-192.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American journal of Applied Mathematics and statistics*, *9*(1), 4-11.

Shultz, K. S., Whitney, D., & Zickar, M. J. (2020). *Measurement theory in action: Case studies and exercises*. Routledge.

Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Simkus, J. (2022). Convenience sampling: Definition, method and examples. *Retrieved Oktober*, *6*, 2022.

Singer, E., & Bossarte, R. M. (2006). Incentives for Survey Participation: When are they "Coercive"? *American Journal of Preventive Medicine*, *31*(5), 411– 418

Skowronek, D., & Duerr, L. (2009). The convenience of nonprobability: Survey strategies for small academic libraries. *College & Research Libraries News*, 70(7), 412-415

Sockeel, P., Dujardin, K., Devos, D., Deneve, C., Destée, A., & Defebvre, L. (2006). The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: validation in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *77*(5), 579-584.

Spitzer, L., & Mueller, S. (2023). Registered report: survey on attitudes and experiences regarding preregistration in psychological research. *PLoS One* 18:e0281086

Stahl, A. C., Tietz, A. S., Dewey, M., & Kendziora, B. (2023). Has the quality of reporting improved since it became mandatory to use the Standards for Reporting Diagnostic Accuracy?. *Insights into Imaging*, *14*(1), 85.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346.

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in psychology*, *8*, 862.

Streiner, D. L., & Kottner, J. (2014). Recommendations for reporting the results of studies of instrument and scale development and testing. *Journal of advanced nursing*, *70*(9), 1970-1979.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). Health Measurement Scales: A Practical Guide to Their Development and Use (5th ed.). Oxford, UK: Oxford University Press

Swan, K., Speyer, R., Scharitzer, M., Farneti, D., Brown, T., Woisard, V., & Cordier, R. (2023). Measuring what matters in healthcare: a practical guide to psychometric principles and instrument development. *Frontiers in Psychology*, *14*, 1225850.

Swank, J. M., & Mullen, P. R. (2017). Evaluating evidence for conceptually related constructs using bivariate correlations. *Measurement and evaluation in counseling and development*, *50*(4), 270-274.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296.

Taylor, J. M., & Radford, E. J. (1986). Psychometric testing as an unfair labour practice. *South African Journal of Psychology*, *16*(3), 79-86.

Teresi, J. A., Yu, X., Stewart, A. L., & Hays, R. D. (2022). Guidelines for designing and evaluating feasibility pilot studies. *Medical care*, *60*(1), 95-103.

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psychol.* 37, 1–2.

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical and Philosophical Psychology, 41*(1), 58–84.

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, 679 – 686.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European review of applied psychology*, *54*(2), 119-135.

Wang, Y. P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. *Revista brasileira de psiquiatria*, *35*(4), 416-431.

White, S. J., Chau, M., Arruzza, E., Ong, M., John, H., Theiss, R., ... & To, M. S. (2025). Assessment of Standards for Reporting of Diagnostic Accuracy (STARD) 2015 guideline adherence in medical imaging diagnostic accuracy studies published in 2023. *Journal of Clinical Epidemiology*, *179*, 111654.

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in psychometrics. *Perspectives on Psychological Science*, *17*(3), 788-804.

Williams, B., FitzGibbon, L., Brady, D., & Christakou, A. (2025). Sample size matters when estimating test–retest reliability of behaviour. *Behavior Research Methods*, *57*(4), 1-25.

Yusoff, M. S. B. (2019). ABC of response process validation and face validity index calculation. *Educ Med J*, *11*(10.21315).