Psychometric Approaches for Analyzing Mobile Psychological Data: Which One is Best?

Marat Assanovich

Oleg Skugarevsky

Mikhail Kaspartov

Andrew Sokol

Published online: November, 2023

SUMMARY. This article explores the application of psychometric methods in mental health assessments conducted through mobile applications. Its primary objective is to identify the most effective psychometric approach for processing mobile psychological data.

The paper conducts a thorough review of three key psychometric measurement models: Classical Test Theory (CTT), Rasch Measurement Theory (RMT), and Item Response Theory (IRT). This evaluation assesses each model's capability in creating and analyzing measurement scales, with a special focus on their adaptability to mobile data collection contexts.

CTT, known for its simplicity, is limited in estimating response patterns, particularly when dealing with polytomous Likert-like items, and lacks robust methods for setting measurement-relevant cut-off criteria. IRT stands out in its analysis of item parameters but is hindered by its inability to support the creation of a measurement scale based on total scores, which is a significant limitation in the context of mobile data. In contrast, RMT provides an extensive range of tools for developing measurement scale models. These tools include capabilities for estimating item parameters, analyzing individual response patterns, and establishing interpretive cut-off criteria. The study ultimately determines that Rasch Measurement Theory is the optimal approach for analyzing psychological data collected via mobile applications, due to its comprehensive methodology and suitability for the unique demands of mobile data analysis.

Key words: mobile psychological assessment, classical test theory, IRT, Rasch measurement theory

Recent years have seen a significant expansion in the use of patient-reported outcomes within healthcare, particularly in psychiatry and clinical psychology. Although some instruments are administered by healthcare professionals, the majority are self-completion tools, extensively used in both clinical practice and research (Arean et al., 2016; Harari et al., 2016; Lecomte et al., 2020). These instruments offer the benefit of reducing the assessment burden on patients and are practical for use in large populations where conducting structured clinical interviews may be challenging or impractical.

There has been a noticeable increase in the number of mobile apps targeting latent constructs such as emotional well-being, mood, anxiety, and depression (Heron & Smyth, 2010; Lecomte et al., 2020; Martín-Martín et al., 2021; Meyerhoff et al., 2021). It is critical to ensure that clinical outcome assessments used in mental health apps are both valid and reliable. Employing tools with untested diagnostic properties can lead to the dissemination of misleading results. Most mobile mental health applications utilize well-established, psychometrically sound self-assessment tools, such as the GAD-7, PHQ-9, and others (Kroenke et al., 2007; Plummer et al., 2016; Sapra et al., 2020). However, when adapted for mobile applications, even tools with strong psychometric properties must be scrutinized for how they are interpreted and used in this new format. This makes assessing the alignment of the data with its intended diagnostic purpose extremely important.

In the context of mobile usage, it is essential that the data accurately capture the intended concept, such as depression or anxiety. Given that data interpretation often relies on total scores, it is crucial that these scores are based on an appropriate measurement model. This ensures the psychometric validity and measurability of the summed scores (Embretson & Reise, 2013; Mari et al., 2021). Three primary measurement models in psychometric data analysis are classical test theory (CTT), Rasch measurement theory (RMT), and item response theory (IRT). This article aims to explore which measurement models are most suitable for mobile psychological data (Cappelleri et al., 2014; Embretson & Reise, 2013; McClimans et al., 2017).

When measuring a construct such as depression, indicators or items related to the construct are identified, ideally aligning with the underlying theoretical framework. The likelihood of an individual endorsing a specific item should correspond with their level of the attribute being measured. For instance, someone with more severe depression is

more likely to endorse an item indicating low interest, compared to someone who is not depressed. While this item alone does not directly measure depression, it contributes to a broader depression score when combined with other related items. The development of a reliable set of items to measure a specific construct adheres to various psychometric theories and requirements, as detailed in existing literature (Mari et al., 2021; McClimans et al., 2017; Mohamad et al., 2015).

To date, most self-reported outcomes questionnaires have been evaluated using Classical Test Theory (CTT) (Cappelleri et al., 2014). CTT posits that an observed indicator (such as a test score) is linked to a latent variable (the 'true score'), with random error influencing the observed score. This concept of linking observed events with latent factors traces back to the pioneering work of Francis Galton, who laid the scientific foundation for this field. Building on the concept of correlation, British psychologist Charles Spearman further developed the classical test theory, or true score model. Between 1904 and 1913, Spearman introduced logical and mathematical arguments to address the presence of errors in test scores, positing that the correlation between 'erroneous' scores is lower than that between their 'true objective values' (Cappelleri et al., 2014; McClimans et al., 2017). Spearman's CTT postulates that any test score (X) can be represented as a composite of a true score (T) and a random error (E), expressed as $X = T + E$.

CTT does not view the true score as the exact number of items a subject can answer correctly or the precise score they can receive. Instead, it conceptualizes a 'true score' as a theoretical construct, equivalent to an ideal measure obtained without error (Embretson & Reise, 2013). In CTT, the true measure of the latent variable is not directly determined; rather, it's the expected equivalence to the theoretical true score that is considered acceptable. The observed test score serves as an indicator of a theoretically existing true score, differing from it by the amount of random error. CTT suggests that with multiple studies using alternative test forms on the same subjects, the average test score will approximate the true score. These alternative test scores are assumed to have a normal distribution and equal variances, symmetrically distributed around the true score. A larger variance in observed scores indicates a greater measurement error. The standard deviation of each subject's random error distribution shows the magnitude of this error. Since all subjects share the same standard deviation, this value alone suffices to estimate the expected error, termed the "standard measurement error" (Cappelleri et al., 2014; de Champlain, 2010; McClimans et al., 2017).

CTT aims to minimize measurement error by maximizing the approximation between observed and true scores. As the true score's measurement error is unknown, approximation from X to T is sought by enhancing test reliability, hence CTT's strong emphasis on test reliability. Within the Classical Test Theory (CTT) framework, reliability is commonly estimated using Cronbach's alpha, which assesses scale reliability through internal consistency. CTT lacks a unified analytical apparatus, instead relying on statistical methods like correlation and factor analysis to validate and verify data reliability. This approach examines the psychometric properties of a test independently, without integrating them into a singular analytical method. Despite these limitations, CTT has been a widely used measurement model throughout the 20th century due to its simplicity, adaptability to various psychological variables and measurement contexts, and ease of calculation for respondent scores. Most data analysis software readily provides basic statistics and reliability coefficients necessary for item and test analysis. Clinicians widely accept CTT's psychometric properties, such as reliability, validity, and sensitivity to change.

However, CTT falls short of the scientific rigor required for high-stakes clinical measurements due to several key limitations: (a) it produces ordinal, not interval, measurements; (b) scores are scale and sample dependent; (c) scale properties like reliability and validity vary with the sample; and (d) while suitable for group-level inferences, CTT cannot accurately measure individual patients. Consequently, test parameters can vary and become unstable across different samples. Furthermore, CTT does not provide methods for detecting falsified responses, particularly in mobile test data, leading to potential inaccuracies and compromised validity. It also lacks criteria for valid and reliable interpretation of mobile data (Cappelleri et al., 2014; de Champlain, 2010; Embretson & Reise, 2013; McClimans et al., 2017). Since the 1930s, Classical Test Theory (CTT) has been foundational in test development. However, recent decades have seen a paradigm shift due to its noted limitations. Increasingly, Item Response Theory (IRT) models and Rasch measurement theory are supplanting CTT in psychological measurements (de Champlain, 2010; McClimans et al., 2017).

IRT represents an alternative approach to psychometric analysis, focusing on the congruence between a chosen model and the structural and statistical patterns of mobile assessment data. This theory describes the functional relationship between a person's response to a test item and a latent variable (e.g., emotional well-being, depression, anxiety), symbolically represented by the Greek letter "theta" (θ). Essentially, individuals

with higher θ levels are more likely to endorse a test item, while those with lower θ are less likely to do so. The core of IRT models lies in statistically analyzing subjects' response patterns to specific items (Cappelleri et al., 2014; de Champlain, 2010; McClimans et al., 2017).

IRT models estimate the relationship between θ and the probability of a particular response. Typically, these relationships are described by a normal ogival or logistic function, with the logistic function being more prevalent and practical in modern psychometrics. IRT models use several parameters to describe the interaction between a person and a test item. The first, theta (θ), characterizes each subject. The item difficulty, denoted as "b", is evaluated similarly to the measured construct. In IRT, item difficulty corresponds to the θ level where 50% of subjects are likely to provide the key response. Additionally, IRT models incorporate an item discrimination parameter, "a", which assesses an item's ability to differentiate between subjects with varying levels of θ. Some models also include a "c" parameter, representing a lower asymptote or guessing parameter, accounting for the likelihood of low θ subjects providing accurate responses.

Depending on the number of item parameters, IRT employs one-parameter (only b), two-parameter (b and a), or three-parameter (b, a, and c) logistic models. These models graphically depict functional relationships between model parameters and response probability via item characteristic curves (ICCs). An ICC, essentially an S-shaped regression line, illustrates how the probability of a key response to a test item varies with θ. The item discrimination (a) is indicated by the ICC curve's slope, which shows how probabilities shift with changes in θ. The item difficulty (b) is graphically represented by the ICC's position relative to the horizontal axis, with a rightward shift indicating a more challenging item (DeMars, 2010; Embretson & Reise, 2013).

Item Response Theory (IRT) models offer several advantages over Classical Test Theory (CTT) models in psychological measurement. A key strength of IRT is its capacity to produce an invariant measurement structure, where person and item parameter estimates are independent of the specific sample from which they are derived. Once measurement parameters are established in one data sample, they can be reliably reproduced in other samples. Additional benefits of IRT include the ability to measure at interval levels and to determine the degree of fit for both items and individuals. There are also notable differences between CTT and IRT in terms of item scaling and local accuracy. CTT assigns equal measurement weight to each item, resulting in uniform

scoring (de Champlain, 2010; McClimans et al., 2017). In contrast, IRT models can differentiate items based on their placement within the measured construct, such as identifying items that signify varying degrees of depression severity. This allows for a more nuanced positioning of respondents on the construct continuum, considering their agreement with specific items that reflect certain symptoms. Furthermore, while CTT employs a constant error across all levels of the measured variable, IRT adapts the measurement error according to variable levels, enhancing overall measurement accuracy. In the realm of mobile data evaluation, IRT models are superior, thanks to their measurement invariance, capability to analyze response patterns, and the differentiated analysis of item and person parameters.

For mobile psychological assessments, IRT models are instrumental in determining item difficulty and discrimination. These parameters aid in assessing the efficacy of test items within mobile data collection contexts. However, the construction of a complete IRT-based psychological measurement is not without challenges. Specifically, IRT does not account for the total test score, a critical factor in interpreting mobile data. The discrimination parameter may lead to varying ability levels for respondents with identical total scores. Therefore, while IRT models offer valuable insights, their utility in analyzing mobile data is somewhat limited (Embretson & Reise, 2013; Mari et al., 2021; Reise & Revicki, n.d.).

The Rasch Measurement Theory (RMT), initially known as the Rasch model (RM) or Rasch analysis, represents a significant approach in unidimensional measurement modeling (Andrich & Marais, 2019; Boone & Staver, 2020). This model calculates the relationship between item difficulty (e.g., the level of depression expressed by an item) and person ability (e.g., an individual's depression level) by evaluating the ratio of positive or negative endorsements and expressing the difference as log-odds. In RMT, a person's likelihood of endorsing an item is logistically related to the difference between their ability level and the item's difficulty.

Although formally similar to a one-parameter logistic IRT model, the Rasch model was developed independently by Danish mathematician Georg Rasch (Andrich & Marais, 2019; Bond et al., 2020). Diverging from other IRT models, RMT uniquely incorporates a total score during parameter estimation, effectively integrating individual response analysis with a summary score. Unlike CTT and IRT, which do not adequately assess the suitability of the total score for measuring latent traits, RMT specifically determines

whether the summary score can be used for data interpretation (Engelhard, 2013; Wilson & Fisher, n.d.).

Rasch analysis refers to a one-parameter model, focusing on the item difficulty parameter. In clinical contexts, this translates to the likelihood of endorsing an item that reflects a specific ability or symptom severity. The model suggests that items representing more severe symptoms (or higher abilities) have lower endorsement odds. The Rasch model's defining characteristic is invariant measurement, asserting that person scores should not depend on the specific items and vice versa (Boone, 2016; Engelhard, 2013). By probabilistically analyzing respondents' answers, RMT evaluates whether response patterns form a scale that meets the requirements of fundamental measurement, including additivity and interval scaling. This approach results in item and person parameters being calculated independently, leading to estimates that are free of sample and test item biases. However, it's necessary to verify empirically that the assumed invariance holds true. When the model fits the data structure, log-odds of item endorsements provide reasonable estimates of a person's underlying ability. Moreover, RMT has an advantage over CTT and IRT as it transforms raw ordinal data into interval variables, using log-odds units as the measure (Andrich & Marais, 2019; Boone & Staver, 2020; Engelhard, 2013; Engelhard Jr. & Wind, 2017; Khine, 2020).

Rasch analysis represents a comprehensive approach to measurement, determining several key aspects: the transformation of item sets and respondent reactions into interval measures; the unidimensionality of items, ensuring they measure a single construct; construct validity, confirming items align with the scale's constructive orientation; the suitability of respondent reactions for ability measurement, distinguishing between genuine and invalid responses; the scale's reliability and separation attributes; the precision of test items in assessing varying ability levels; and the appropriateness of total scores for assessment and data interpretation (Boone, 2016; Boone et al., 2014; Boone & Staver, 2020; da Rocha et al., 2013; Engelhard, 2013; Engelhard Jr. & Wind, 2017; Green & Frantom, 2002).

A diverse array of models within Rasch Measurement Theory is available for analysis, tailored to the format of test items. The classical Rasch model, typically employed for dichotomous items offering binary 'yes' or 'no' responses, is a standard approach. For Likert-like items, two polytomous models are employed: the Rating Scale

Model and the Partial Credit Model (Boone & Staver, 2020; da Rocha et al., 2013; Khine, 2020).

The Rasch Measurement Theory offers several components crucial for evaluating mobile assessment results.

Item Difficulty Analysis. This pertains to the level of ability an item measures. Analyzing item difficulty in mobile assessments helps arrange test items hierarchically based on their difficulty. Since item difficulty indicates the likelihood of an item being endorsed relative to a person's ability, it reflects the frequency of the symptom's occurrence in the population. Common symptoms have low difficulty, whereas rare symptoms exhibit high difficulty. This analysis is vital for gauging the effectiveness of a mobile assessment questionnaire in accurately measuring abilities at various severity levels (Boone & Staver, 2020; McClimans et al., 2017; Mohamad et al., 2015).

Category Thresholds Analysis. Commonly, mobile psychological assessments use Likert-like items with multiple response categories. Ideally, these categories should form a continuum reflecting increasing levels of a trait, such as depression. For instance, "several days" should indicate a lower level of depression than "more than half of the days." Disordered rating scale categories can imply misunderstandings in how the scale is interpreted by participants and researchers. Rasch analysis allows for the determination of category threshold parameters – points between adjacent response categories where both are equally likely. The ordering of threshold parameter values should reflect the sequence of answer categories, ensuring the highest probability of each answer within its category interval. Disordered thresholds, which may arise from ambiguous wording or respondents' inability to distinguish between options, might necessitate item rescoring (Boone, 2016; Boone & Staver, 2020; Khine, 2020).

Fit Statistics. To conform to Rasch model properties, empirical data must align with model predictions. Fit statistics evaluate the data's conformity to the model and the usefulness of the measure scale. These include average fit (mean square and standardized) of persons and items, and fit statistics related to the appropriateness of rating scale categories. Fit statistics are calculated by squaring the differences between observed and expected responses, averaging these across pairs, and standardizing them. Ideal fit produces mean square and standardized fit indices of 1.0 and 0.0, respectively. Person fit statistics, unique to the Rasch model, assess respondent consistency, identifying potential issues like inattention or confusion. They help determine

the reliability of data in a mobile dataset for further analysis. Item fit statistics, on the other hand, check if items appropriately measure the intended ability. Misfitted items, possibly due to complexity or measuring a different construct, are flagged for review to ensure constructive validity (Boone, 2016; Boone et al., 2014; Green & Frantom, 2002; Wright & Masters, 1990).

Item-Person Map (Targeting). A key aspect of Rasch analysis is aligning item difficulty with a person's ability. Accurate measurement requires differentiation across the full spectrum of abilities, necessitating a range of item difficulties. Rasch analysis uses a person-item map to illustrate item targeting relative to patient abilities and calculates the mean differences between items and patients. Optimal targeting occurs when the average difficulty of items matches the average ability of persons. A larger mean difference signifies poorer targeting (Boone et al., 2014; Green & Frantom, 2002).

Unidimensionality Analysis. Essential for the structural validation of psychometric scales, unidimensionality means only one latent dimension (like depression) contributes to the common variance. For scales assessing a single ability, all items should measure just that. If a scale is unidimensional, local independence is observed, meaning the latent dimension explains item relationships, and no significant intercorrelation exists between items. Rasch analysis, despite modeling one-dimensional measurement scales, can detect multidimensionality more sensitively than traditional factor analysis (FA), using principal components analysis of standardized residuals. Unidimensionality is assessed by the eigenvalue of the first contrast in the principal components matrix, with a value of 2 or less indicating a unidimensional scale (Boone & Staver, 2020; Engelhard, 2013).

Differential Item Functioning (DIF). In Rasch analysis, items should measure the same trait consistently across different groups. DIF occurs when an item's estimates vary among groups. It suggests the presence of construct-irrelevant variance affecting the scores. DIF analysis, crucial in mobile dataset research due to diverse populations, assesses measurement invariance across different groups based on variables like pathology, gender, age, education level, and country. This analysis often employs the Mantel-Haenszel procedure, comparing item responses between reference and focal groups with equivalent attribute levels (Boone, 2016; Boone & Staver, 2020; da Rocha et al., 2013).

Interval Scale, Logits. The Rasch model transforms raw scores into logit interval measures. Logits, theoretically ranging from negative to positive infinity, typically span from -5 to +5 logits in practical applications. Each raw total score corresponds to a specific log-odd and has an equivalent logit value, complete with a standard error of estimate. This logit scale accurately reflects the measurement differences between raw total scores across ability continuums (Engelhard, 2013; Engelhard Jr. & Wind, 2017; Wilson & Fisher, n.d.).

Rasch Change Index (RCI). The Standard Error (SE) associated with each logit value, which corresponds to a raw score, determines the precision of the measurement. A significant difference between two logit measures can be calculated using their respective standard errors. Since there is an equivalent relationship between logits and raw scores, a significant difference in logits implies a corresponding measurement difference in raw scores. Essentially, the RCI represents the statistically significant difference between two logits and the significant measurement distance between the raw total scores corresponding to these logits. An RCI value is calculated based on the standard error of the difference between logits and the Z-value, which represents the selected probability. Commonly, RCI is computed using probabilities of 95% (Z=1.96) and 90% (Z=1.645). The difference between logits is considered statistically significant if the RCI value exceeds the arithmetic difference between them. By applying RCI consistently from the starting to the ending logit, we can segment the entire scale into statistically significant ranges reflecting different levels of ability. This method allows for scientifically justified cut-off criteria for the interpretation of mobile psychological and clinical outcome assessments (Caronni et al., 2021).

Reliability and Separation Statistics. Rasch analysis estimates reliability for both items and persons. Person reliability is analogous to the classic Cronbach's alpha test reliability and indicates the reproducibility of a person's ability order for a specific set of items across a sample. Item reliability measures the consistency of item difficulty ordering across respondents for a set of items. Separation statistics use standard error units to quantify the dispersion of both items and persons. In essence, it represents the number of distinct levels assignable to the sample of items and individuals. For an instrument to be effective, its separation should exceed 1, with higher values indicating a broader spread of items and persons. Estimating separation is crucial, especially in the context of mobile assessment data. Low separation statistics suggest that the mobile data sample

does not adequately fit the items, or vice versa (Boone & Staver, 2020; Green & Frantom, 2002; Khine, 2020).

In this study, we explored the feasibility and scientific validity of mobile psychological assessments. Three key approaches emerged for ensuring scientific robustness in data: Classical Test Theory (CTT), Item Response Theory (IRT), and Rasch Measurement Theory (RMT). While CTT offers a simpler method for data analysis, it falls short in estimating response patterns in relation to test items and is limited when handling polytomous Likert-like items. Additionally, CTT lacks methods to establish measurement-relevant cut-off criteria, compromising reproducibility and objectivity in mobile data analysis. In contrast, IRT excels in analyzing mobile assessment data, particularly in terms of item parameters. However, it does not support the creation of a measurement scale based on total scores and might necessitate item discrimination, which is critical for mobile data. RMT provides comprehensive tools for developing measurement scale models from mobile datasets. These tools range from estimating item parameters and individual response patterns to establishing interpretive cut-off criteria. Therefore, for this study, Rasch Measurement Theory emerges as the most suitable psychometric approach for analyzing psychological data collected via mobile apps.

References

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Arean, P. A., Hallgren, K. A., Jordan, J. T., Gazzaley, A., Atkins, D. C., Heagerty, P. J., & Anguera, J. A. (2016). The use and effectiveness of mobile apps for depression: Results from a fully remote clinical trial. *Journal of Medical Internet Research*, *18*(12). https://doi.org/10.2196/JMIR.6482/JMIR_V18I12E330_APP2_PDF.PDF

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model : fundamental measurement in the human sciences*. Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development. *Life Sciences Educ.*, *15*(4), 1–7.

Boone, W. J., & Staver, J. R. (2020). Advances in Rasch Analyses in the Human Sciences. *Advances in Rasch Analyses in the Human Sciences*. https://doi.org/10.1007/978-3-030-43420-5

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch Analysis in the Human Sciences. In *Book*. https://doi.org/10.1007/978-94-007-6857-4

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Caronni, A., Picardi, M., Gilardone, G., & Corbo, M. (2021). The McNemar Change Index worked better than the Minimal Detectable Change in demonstrating the change at a single subject level. *Journal of Clinical Epidemiology*, *131*, 79–88. https://doi.org/10.1016/j.jclinepi.2020.11.015

da Rocha, N. S., Chachamovich, E., de Almeida Fleck, M. P., & Tennant, A. (2013). An introduction to Rasch analysis for Psychiatric practice and research. *Journal of Psychiatric Research*, *47*(2), 141–148. https://doi.org/10.1016/j.jpsychires.2012.09.014

de Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. In *Medical Education* (Vol. 44, Issue 1, pp. 109–117). https://doi.org/10.1111/j.1365-2923.2009.03425.x

DeMars, Christine. (2010). *Item response theory*. Oxford University Press.

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. In *Item Response Theory for Psychologists*. Taylor and Francis. https://doi.org/10.4324/9781410605269

Engelhard, G. (2013). *Invariant measurement : using Rasch models in the social, behavioral, and health sciences*. Routledge. https://books.google.com/books/about/Invariant_Measurement.html?hl=ru&id=ZRkK7RFsIOkC

Engelhard Jr., George., & Wind, Stefanie. (2017). *Invariant Measurement with Raters and Rating Scales : Rasch Models for Rater-Mediated Assessments*. Taylor and Francis. https://books.google.com/books/about/Invariant_Measurement_with_Raters_and_Ra.html?hl=ru&id=izZDDwAAQBAJ

Green, K., & Frantom, C. (2002). Survey development and validation with the Rasch model. *International Conference on Questionnaire Development, Evaluation, and Testing*, 1–30.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*(1), 1–39. https://doi.org/10.1348/135910709X466063

Khine, M. S. (2020). Rasch measurement: Applications in quantitative educational research. *Rasch Measurement: Applications in Quantitative Educational Research*, 1–281. https://doi.org/10.1007/978-981-15-1800-3/COVER

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Lö, B. (2007). *Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection*. www.annals.org

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Lecomte, T., Potvin, S., Corbière, M., Guay, S., Samson, C., Cloutier, B., Francoeur, A., Pennou, A., & Khazaal, Y. (2020). Mobile apps for mental health issues: Meta-review of meta-analyses. *JMIR MHealth and UHealth*, *8*(5). https://doi.org/10.2196/17458

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the Sciences*. https://doi.org/10.1007/978-3-030-65558-7

Martín-Martín, J., Muro-Culebras, A., Roldán-Jiménez, C., Escriche-Escuder, A., De-Torres, I., González-Sánchez, M., Ruiz-Muñoz, M., Mayoral-Cleries, F., Biró, A., Tang, W., Nikolova, B., Salvatore, A., & Cuesta-Vargas, A. (2021). Evaluation of android and apple store depression applications based on mobile application rating scale. *International Journal of Environmental Research and Public Health*, *18*(23). https://doi.org/10.3390/IJERPH182312505

McClimans, L., Browne, J., & Cano, S. (2017). Clinical outcome measurement: Models, theory, psychometrics and practice. *Studies in History and Philosophy of Science Part A*, *65–66*, 67–73. https://doi.org/10.1016/j.shpsa.2017.06.004

Meyerhoff, J., Liu, T., Kording, K. P., Ungar, L. H., Kaiser, S. M., Karr, C. J., & Mohr, D. C. (2021). Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: Longitudinal cohort study. *Journal of Medical Internet Research*, *23*(9). https://doi.org/10.2196/22844

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the Validity and Reliability of Research Instruments. *Procedia - Social and Behavioral Sciences*, *204*, 164–171. https://doi.org/10.1016/j.sbspro.2015.08.129

Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016). Screening for anxiety disorders with the GAD-7 and GAD-2: A systematic review and diagnostic metaanalysis. *General Hospital Psychiatry*, *39*, 24–31. https://doi.org/10.1016/j.genhosppsych.2015.11.005

Reise, S. Paul., & Revicki, D. A. (n.d.). *Handbook of item response theory modeling : applications to typical performance assessment*. 465. Retrieved September 15, 2022, from https://books.google.com/books/about/Handbook_of_Item_Response_Theory_Modelin.html?hl=ru&id=yDiLBQAAQBAJ

Sapra, A., Bhandari, P., Sharma, S., Chanpura, T., & Lopp, L. (2020). Using Generalized Anxiety Disorder-2 (GAD-2) and GAD-7 in a Primary Care Setting. *Cureus*, *12*(5). https://doi.org/10.7759/CUREUS.8224

Wilson, M., & Fisher, W. P. (n.d.). *Psychological and social measurement : the career and contributions of Benjamin D. Wright*. Retrieved September 15, 2022, from https://books.google.com/books/about/Psychological_and_Social_Measurement.html?hl=ru&id=ejxEDwAAQBAJ

Wright, B. D., & Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions*, *3*(4), 84–85.