The European Journal of Education and Applied Psychology 2024, No 1
ISSN 2310-5704

PREMIER Publishing
ppublishing.org

Section 6. Practical psychology

DOI:10.29013/EJEAP-24-1-40-46

PSYCHOMETRIC CREATION OF ABILITY SCALES: THE METHODOLOGY OF CALIBRATION AND STANDARDIZATION WITHIN THE FRAMEWORK OF THE RASCH MODEL

Tuzhilkina Irina [1]

[1] Head of the Sector for Development and Support of Testing in Educational Programs, Moscow City University of Management of the Government of Moscow named after Yu. M. Luzhkov, Russia, Moscow

Cite: *Tuzhilkina, I. (2024). Psychometric creation of ability scales: the methodology of calibration and standardization within the framework of the Rasch model. European Journal of Education and Applied Psychology 2024, No 1.* https://doi.org/10.29013/EJEAP-24-1-40-46

Abstract

This article reveals the theoretical and practical aspects of the psychometric construction of ability scales based on the Rasch model. This approach allows us to look at the measurement of abilities as a process of evaluating a hidden trait through the harmonious interaction of task parameters and characteristics of subjects on a single logit scale. The theoretical foundations of constructing ability scales are considered in detail, including the requirements for their one-dimensionality, local independence, and compliance with empirical data from the measurement model. The key stages of designing the scale are described: determining the measured property, creating a task bank, conducting pilot tests, analyzing the data structure, calibrating tasks, and evaluating the parameters of the subjects. Special attention is paid to the procedures for standardizing the scale, converting initial scores into linear measures, and quality control of the scale using indicators of reliability, validity, and measurement invariance. The paper shows how using the Rasch model helps to identify complex tasks, assess the scale's compliance with the level of the sample under study, and make the results more reliable. Some practical difficulties of using this model in the process of creating ability scales were also considered, and prospects for its further use in psychometric research were identified.

Keywords: *psychometry, Rasch model, ability scale, psychometric design, task calibration, standardization, reliability, validity, measurement invariance, latent trait.*

Relevance of the study

The relevance of the topic is explained by the fact that in modern psychodiagnostics and educational measurements, high accuracy, comparability, and validity of ability assessment results are extremely important.

The Rasch model is one of the most developed psychometric approaches, as it allows us to assess the level of abilities of subjects and the complexity of tasks on a single logit scale. In addition, the model allows you to verify the compliance of empirical data with the requirements of the measurement model. This makes it particularly valuable for creating scales that require calibration of tasks, identification of ineffective test elements, and obtaining more stable measurement results.

Additional relevance to the topic is given by the fact that in the scientific literature, the Rasch model is actively used to create, validate, and standardize measurement tools in psychology, pedagogy, and other related fields. As review publications and applied research show, interest in Rasch modeling and its use for psychometric scale verification is constantly growing. This indicates that the calibration and standardization methodology based on this model is an important tool for developing reliable aptitude assessment tools.

Russian research also highlights the importance of psychometric analysis and standardization in the creation of diagnostic techniques. Without these procedures, it is difficult to interpret the results and apply them to different samples. Therefore, the study of the psychometric construction of ability scales based on the Rasch model is of interest both from a theoretical and practical point of view.

The purpose of the study

The purpose of this study is to examine the methodological principles of creating psychometric scales of abilities based on the Rasch model. We will also look at the calibration, standardization, and quality control procedures of the scale, which are key to developing a reliable and sustainable measuring instrument.

Materials and research methods

Published scientific papers on the application of the Rasch model in various fields were used as sources for the study: psychometry, psychology, pedagogy, and related disciplines. In addition, the paper considered examples of empirical data used to assess the quality of scales.

The research methodology included the analysis of scientific literature, comparative generalization of theoretical concepts, and interpretation of the results of practical research related to the calibration of tasks, assessment of subject characteristics, standardization of scales, and control of their psychometric properties.

The results of the study

The construction of psychometric ability scales is based on theoretical concepts related to the measurement of latent characteristics that cannot be observed directly but can be assessed based on task results. In modern psychometry, such scales are not just a sum of points but a carefully designed system based on models. This system verifies important aspects such as the unidimensionality of the construct, the independence of responses, and the consistency of tasks (Maslak, A.A., 2021, p. 115). For an ability scale, it is critically important that the results allow subjects to be sorted by the level of ability development and tasks by the level of difficulty. Only in this case does the measurement become interpretable and suitable for comparison. It should be noted that it is precisely the arrangement of subjects and tasks on the general scale that opens the way from ordinal observations to a more rigorous measurement description of the results.

The Rasch model is a fundamental basis for the development of such scales, as it establishes a clear relationship between the level of abilities of the subject and the complexity of the task. In the dichotomous version of the model, the probability of a correct answer is determined by the difference between ability and difficulty, and the difficulty value is interpreted as the point at which the probability of successful completion of the task is 0.5. This allows us to evaluate both parameters on a single logistic scale and check whether the empirical data meet the requirements of the model. Unlike other approaches that only assess the consistency of data with the sample as a whole, the Rasch approach considers the model as a normative basis for measurement. It is not the model that adapts to any data; rather, the tasks and answers must demonstrate acceptable compliance with the measurement design.

To construct a scale of abilities within the Rasch model, there are three key procedures:

calibrating the difficulty of tasks, assessing the ability level of individuals, and analyzing the correspondence of data with the model.

It is significant that real-world Rush analyses allow not only to obtain an overall assessment of the reliability of the scale, but also to identify specific problematic tasks. In the study of the perceived stress scale for a sample of 752 people, one of the tasks in the multidimensional model clearly exceeded the acceptable range: for item 8, the values of outfit and infit were 1.555 and 1.535 in the PSS-13 version, and in the PSS-10 version – 1.576 and 1.568; while for most other tasks, the indicators remained close to 1.0. In the same paper, it is reported that the reliability of measurements in the one-dimensional model was 0.799 for PSS-13 and 0.718 for PSS-10, which shows the practical value of Rasch calibration: it allows you not to limit yourself to the total score, but to check the quality of each element of the scale (Table 1).

Table 1. *An example of real indicators of compliance with the tasks of the Rasch model*

Scale and model	Task	Difficulty parameter	RMS outfit	RMS infit	Interpretation
PSS-13, a multidimensional credit model	8	–0.033	1.555	1.535	Marked Nonconformity
PSS-10, a multidimensional credit model	8	–0.026	1.576	1.568	Marked Nonconformity
PSS-10, a rating scale model	6	–0.232	1.396	1.406	Borderline Deviation
PSS-13, a rating scale model	2	0.233	1.081	1.088	Acceptable Conformity
PSS-13, a rating scale model	1	–0.065	0.924	0.924	Acceptable Conformity

A source: (Boluarte-Carbajal, A., 2023).

The process of creating an ability scale in accordance with the Rasch model begins with a clear definition of the measured property and preparation of the content of the future instrument. At this stage, it is important to clearly understand which ability is to be evaluated, form a set of tasks, and determine the format of the answers. This is followed by a pilot study stage, during which primary data is collected and it is checked whether the tasks correspond to the stated content and allow the subjects to be distinguished by the level of ability development. After that, it is necessary to analyze the structure of the received data to make sure that the tasks really work as a single system for measurement. Only then can you start creating the scale itself.

The next stage is to check the effectiveness of the tasks within the framework of the selected model. At this stage, the compliance of the answers with the requirements of the Rasch model is assessed, tasks that give unstable results are identified, and it is analyzed whether the integrity of the measured feature is violated. If the scale has several gradations, then it is checked whether the subjects can distinguish between neighboring categories of responses (distractors). If necessary, individual tasks are excluded or redistributed into content blocks.

The process of calibrating tasks and evaluating participants' abilities is a fundamental step in building a scale of individual differences. According to the Rasch model, the complexity of tasks and the skill level of subjects are calculated on a single logarithmic scale, which makes it possible to compare these two parameters within a single measurement. This allows you to identify tasks that are easier and more difficult, as well as assess how well a specific set of tasks corresponds to the level of the sample under study. If the tasks turn out to be too simple or too difficult for most participants, this may indicate the need to refine the scale, as its measurement capabilities may be limited.

The assessment of the subjects' parameters allows us to place them on a scale of abilities. The calibration of tasks helps to determine their difficulty level. At the same time, it is important to analyze how well the empirical data fit the model, as this shows which tasks may introduce errors into the measurement process. If individual items have unstable indicators, they should be reviewed. Calibration in the Rasch model helps to create a scale where the tasks and subjects' results are connected by a consistent measurement logic. This makes the scale more reliable, stable, and suitable for standardization in future studies.

The standardization of ability scales within the framework of the Rasch model involves the transformation of initial ordinal estimates into linear indicators expressed in logits. After that, conversion tables and interpretative scales are created. A review of the application of the Rasch model to rehabilitation measurements has been conducted, and it states that dichotomous and polytomic variants of the model can transform ordinal scales into interval measures if the data meet the requirements of the model. This provides a solid basis for standardization, as the researcher receives not only an overall score, but also a metric-ordered scale that can be used to compare results between subjects and groups (Tennant, A., 2023).

The practical importance of standardization can be clearly seen on the example of the UL–LIMOS scale. A study conducted in 2023 on a sample of 407 patients converted the initial range from 0 to 20 points into logits and then increased it from 0 to 100. The following indicators were obtained for this scale: the reliability of the separation of subjects was 0.90, the lower margin of error was 2.7%, the upper margin of error was 13%, and the average position of subjects on the scale was 1.3 logits. These data are used in the standardization process to assess how well the scale reflects the actual range of abilities and whether there is a significant concentration of results in the lower or upper part of the scale (Van de Winckel, A., 2023).

The figure below shows a map of the distribution of subjects and tasks on the general Rasch scale. Maps of this type, on the left, reflect the distribution of study participants by level of ability, and on the right, the distribution of tasks by degree of difficulty. This tool is used in the standardization process to evaluate the scale targeting, that is, how much the difficulty of tasks corresponds to the level of the sample under study.

The quality control of a scale within the Rasch model involves checking its reliability, validity, and invariance. This includes analyzing model compliance, assessing the reliability and distinctiveness of the scale, verifying local independence, one-dimensionality, and differential functioning of items. Reliability indicates the stability of the differences between individuals, validity ensures that the scale measures the intended construct, and invariance ensures there is no systematic bias in the item parameters across groups.

It is convenient to consider quality control indicators based on specific empirical examples. In a study of the Functionality Appreciation Scale on a sample of 567 adults, the one-dimensionality of the scale and the absence of differential functioning of tasks were confirmed. However, the upper marginal effect was 28.04% and the average position of the subjects reached 3.06 ± 2.07 logits. The authors concluded that the scale, in its current form, is too light for this sample and requires tasks that are more difficult. This example shows that, even with an acceptable structure, a scale may need improvement in terms of matching the level of subjects (Feng, S., 2023).

In the Berg Balance Scale study, designed for people with multiple sclerosis, quality control was expanded by checking external validity. The authors found statistically significant associations with external criteria: the correlation with the ABC scale was 0.523, and with the EDSS scale –0.573. In addition, the internal constructive validity and reliability of the instrument were confirmed. However, the scale turned out to be slightly shifted towards more severe cases, which means that it did not fully correspond to the entire sample in terms of difficulty. These results demonstrate that quality control in the Rasch model includes not only internal verification of the model's compliance but also an assessment of the interpretability of the results based on external criteria (Caselli, S., 2023).

Figure 1. *The distribution map of subjects and tasks on the general Rasch scale (Stelmack, J., 2004).*

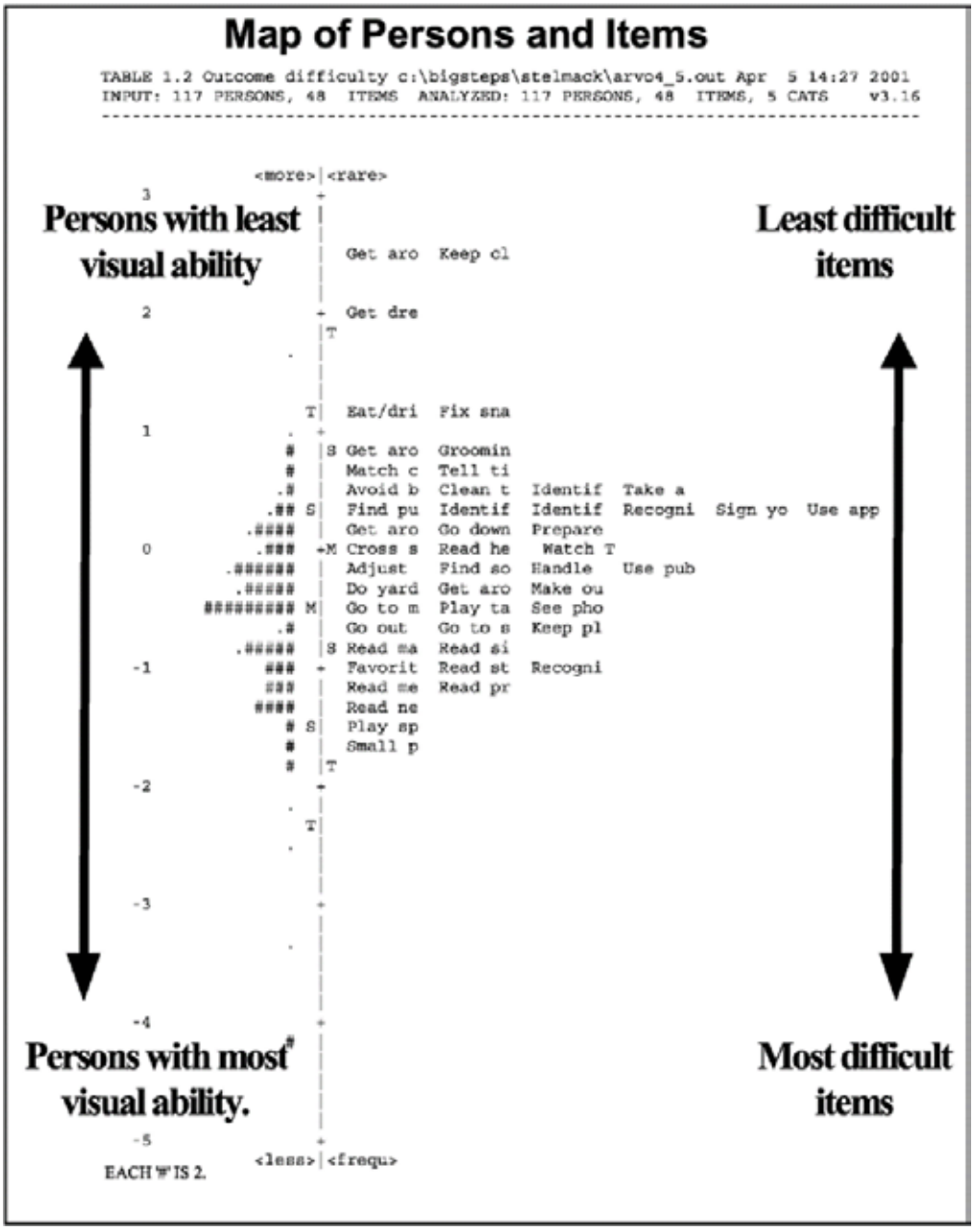

Table 2 shows the practical difficulties and limitations that can be encountered when using the Rasch model to create ability scales. Turning to the applied aspects of the model, it should be noted that its high methodological rigor is both an advantage and a source of some difficulties for researchers. In practice, the scale developer is faced not only with mathematical data processing but also with requirements for the quality of tasks, the structure of the measured construct, and the characteristics of the sample. This requires careful observance of the conditions of application of the model and caution in interpreting the results obtained.

The prospects of using the Rasch model in psychometric research are primarily related to the further development and improvement of measuring tools in psychology, pedagogy, medicine, and other related fields (Efremova, N.F., 2016, p. 68). This approach opens up opportunities not only to verify the quality of existing scales but also to create new, more accurate and comparable tools. In addition, the Rasch model allows you to monitor the work of individual tasks, which makes the

measurement process more efficient. Modern reviews note that the Rasch model retains its importance as a methodology of fundamental measurement and is actively used to improve the quality and accuracy of psychometric scales.

Table 2. *Practical difficulties and limitations of using the Rasch model in developing ability scales*

Practical difficulty	The content of the problem
The requirement of one-dimensionality	The scale should measure a single latent feature. If several properties are combined, the accuracy of the measurement decreases.
Task quality	Unclear or different assignments reduce the quality of the scale and make it difficult to interpret the results.
Tasks that do not match the selection	Tasks that are too light or too complex reduce the effectiveness of the scale.
Selection Restrictions	A small sample size can increase the standard errors and reduce the stability of estimates.
Inconsistency of individual tasks in the model	Some tasks may introduce distortions into the measurement process and require revision or exclusion.
The difficulty of interpreting the results	The logit scale needs to be translated into a more understandable format for practical use.

A source: author's development

An important aspect is the adaptation of scales for different language, age, and social groups. Rasch analysis provides an opportunity to test the stability of measurements and detect task bias in different samples. This makes the model particularly useful for cross-cultural adaptation of methods, the creation of digital tests, and comparative studies where a unified and stable assessment system is required.

Another perspective is to expand the use of the Rasch model in research where it is necessary to evaluate not only abilities but also competencies, attitudes, quality of life, clinical conditions, and educational outcomes.

Conclusions

The conducted analysis allows us to conclude that the Rasch model is an effective tool for developing psychometric scales of abilities. Using it allows you to move from simple scoring to a more rigorous measurement approach that takes into account both the complexity of the tasks and the level of abilities of the subjects. Using the Rasch model, tasks can be calibrated, elements that violate the scale structure can be identified, initial results can be converted into linear measures, and data can be more comparable. The quality of the scale is determined not only by its internal consistency but also by indicators of reliability, validity, invariance, and compliance with the level of the sample under study. Despite some practical limitations related to the need to ensure one-dimensionality, high-quality assignments, and compliance with sample characteristics, the Rasch model still plays an important role in the development, verification, and improvement of psychometric tools used in modern research.

References

Efremova, N.F. (2016). Standardization as a Condition for Ensuring the Quality of University Assessment Tools // International Journal of Applied and Fundamental Research. – No. 2–1. – Pp. 66–70.

Maslak, A.A. (2021). Comparative Analysis of Competency Measurement Methods // New Technologies for Assessing the Quality of Education: Collection of Materials from the

XVI Forum of the Guild of Experts in Professional Education through Online Conferences. – Pp. 115–121.
Boluarte-Carbajal, A. (2023). Psychometric Review of the Perceived Stress Scale under CFA and Rasch Models in Lima, Peru / A. Boluarte-Carbajal, M. Salazar-Conde, S. Alata Vasquez, A. Zegarra-López // Frontiers in Psychology. – Vol. 14. – DOI 10.3389/fpsyg.2023.1160466.
Caselli, S. (2023). When "good" is not good enough: a retrospective Rasch analysis study of the Berg Balance Scale for persons with Multiple Sclerosis / S. Caselli, L. Sabattini, D. Cattaneo [et al.] // Frontiers in Neurology. – Vol. 14. – DOI 10.3389/fneur.2023.1171163.
Feng, S. (2023). Finding functionality: Rasch analysis of the Functionality Appreciation Scale in community-dwelling adults in the US / S. Feng, S. McDaniel, A. Van de Winckel // Frontiers in Rehabilitation Sciences. – Vol. 4. – DOI 10.3389/fresc.2023.1222892.
Stelmack, J. (2004). Use of Rasch person-item map in exploratory data analysis: A clinical perspective / J. Stelmack, J. P. Szlyk, T. Stelmack [et al.] // Journal of Rehabilitation Research and Development. – Vol. 41, no. 2. – pp. 233–241. – DOI 10.1682/JRRD.2004.02.0233.
Tennant, A. (2023). Application of the Rasch measurement model in rehabilitation research and practice: early developments, current practice, and future challenges / A. Tennant, A. A. Küçükdeveci // Frontiers in Rehabilitation Sciences. – Vol. 4. – DOI 10.3389/fresc.2023.1208670.
Van de Winckel, A. (2023). Rasch validation of a new scale to measure dependency in arm use in daily life: the Upper Limb Lucerne ICF-based Multidisciplinary Observation Scale / A. Van de Winckel, B. Ottiger, J. M. Veerbeek [et al.] // Frontiers in Neurology. – Vol. 14. – DOI 10.3389/fneur.2023.1154322.

submitted 02.04.2024;
accepted for publication 16.04.2024;
published 26.04.2024

Contact: irinadezcom@gmail.com