Adam Sagan
Krakow University of Economics (Poland, 0000-0002-5608-179X)

IRT Reliability of Coffee Ritual Behavior Scale - a New Metrics

DOI: 10.69102/CTMJ.2025.020201

ABSTRACT
This study examined the psychometric properties of the Ritual Coffee Consumption Scale using the Rasch model. Results demonstrated good item and person fit, confirming unidimensionality and model conformity. However, the Person Separation Index (PSI ≈ 1.0) and an undefined Item Separation Index (ISI) indicated limited differentiation among respondents and items. A newly developed $Quality_{(i)}$ index, integrating reliability and model fit, revealed the highest precision around the midrange of the latent continuum and reduced accuracy at the extremes. Findings highlight the scale's validity for assessing moderate ritual engagement and suggest extending item difficulty range to enhance measurement coverage and reliability.

Keywords: ritual coffee consumption, IRT model, scale reliability and quality
JEL Classification: M31

Paper received: 31 October 2025 • Paper revised: 21 November 2025 • Paper accepted: 28 November 2025

Introduction

Ritualized consumption of coffee and tea is deeply embedded within the social, cultural, and spiritual lives of numerous societies across the globe. These practices extend far beyond simple acts of consumption, encompassing rich layers of symbolism, interpersonal connection, and communal meaning (Douglas, 1987; Warde, 2005). Whether manifested through the meditative precision of the Japanese tea ceremony or the convivial gatherings over Turkish coffee, each culture has transformed the act of drinking into a ritual that reinforces identity, belonging, and social cohesion (Ratcliffe et al., 2019).

Coffee rituals exhibit remarkable diversity and cultural specificity worldwide. In Ethiopia, widely regarded as the birthplace of coffee, the traditional coffee ceremony occupies a central role in social gatherings, symbolizing hospitality, respect, and community. The process—roasting green beans, grinding them by hand, and brewing the coffee in a jebena (a traditional clay pot)—is both sensory and spiritual, often accompanied by incense, song, and communal sharing. This elaborate ceremony may

last several hours and functions as a vital site of social interaction and cultural continuity (Kucukkomurler and Ozgen 2009).

In Italy, the espresso culture epitomizes efficiency, community, and aesthetic appreciation. Espresso is consumed multiple times daily, particularly in the morning and after meals, often while standing at a café bar rather than sitting. Far from being a solitary act, this practice forms an essential element of the Italian dolce vita ethos, fostering social interaction and reinforcing local community bonds (Morris, 2013).

Turkish coffee, by contrast, represents a slow, deliberate ritual emphasizing preparation and shared experience. Brewed by boiling finely ground coffee with water and sugar in a small pot (*cezve*), it is served unfiltered, producing a dense, sedimented texture at the bottom of the cup. The serving is accompanied by a glass of water and often a piece of Turkish delight, underscoring hospitality and refinement (Hattox, 2014). The ritual also extends to fortune-telling from coffee grounds (*fal*), intertwining sociality, aesthetics, and spirituality.

In France, the café culture is intimately connected to leisure, contemplation, and intellectual exchange. The consumption of coffee—often in the form of café au lait—is embedded in the rhythm of everyday life and public sociability. French cafés historically function as "third places" (Oldenburg, 1999), offering spaces for reading, conversation, and observation of urban life. Coffee drinking thus embodies both individual reflection and collective engagement within French cultural life.

In Poland, coffee and tea rituals occupy a prominent place in everyday social practice, emphasizing hospitality, relaxation, and social connectedness. Offering coffee to guests is a culturally normative gesture of welcome and warmth, reflecting traditional values of hospitality. Coffee is commonly served during family gatherings, holidays, and informal social visits, typically accompanied by sweets or desserts (Olejniczuk-Merta, 2024).

Polish café culture, which flourished in cities such as Kraków, Warsaw and Wrocław during the interwar period, remains integral to urban social life. Coffeehouses historically served as centers for intellectual discourse, artistic collaboration, and political debate. In contemporary Poland, cafés continue to function as multifunctional social spaces—venues for work, study, conversation, and relaxation. Popular styles of coffee include brewed (kawa parzona), espresso, and cappuccino, while instant coffee remains common in domestic settings (Maciejewski et al., 2020).

The ritual of the coffee break - whether at work or in informal gatherings - plays a key role in fostering social connection and psychological rejuvenation. These pauses serve not only physiological, but also symbolic purposes, reaffirming interpersonal bonds and providing a shared sense of rhythm in everyday life (Warde and Martens, 2000). Polish coffee rituals serve multiple functions: hospitality (signaling respect and warmth), socialization (fostering interpersonal relationships and maintaining social networks) and relaxation (moments of leisure and marking transitions between work and private life).

In summary, Polish coffee and tea rituals exemplify how everyday consumption can embody cultural values, social norms, and collective identity. Whether expressed through domestic hospitality or urban café culture, these practices remain vital expressions of sociability and belonging in Polish society.

Theories of ritual consumption

Ritual consumption constitutes a core concept within cultural anthropology and consumer behavior, particularly within the domain of Consumer Culture Theory (CCT) (Arnould & Thompson, 2005). It refers to patterned and symbolically meaningful consumption activities that transcend utilitarian needs and express cultural, emotional, and social values. The following theoretical perspectives provide complementary explanations for ritual consumption.

Ritual consumption theory

Ritual consumption follows socially and symbolically structured patterns rather than merely satisfying functional needs. According to this view, consumption rituals involve prescribed behaviors, objects, and symbols that structure everyday experiences and reinforce collective meanings (Rook, 1985). Ritualized consumption allows individuals to express identity, manage emotions, and sustain social ties. The emphasis lies on the process—the preparation, the context, and the symbolic interactions that accompany consumption. Drinking coffee at a particular time of day, using specific equipment, such as a stovetop espresso maker, and doing so in a particular setting (e.g., with colleagues or friends) exemplifies ritualized consumption in everyday life.

Consumption as social practice theory

Consumption is a socially embedded practice rather than an isolated individual act. Grounded in practice theory, this perspective views consumption as part of socially shared routines and cultural norms (Warde, 2005; Schatzki, 1996). Ritual consumption contributes to the reproduction of social practices that bind individuals to communities and traditions. Through shared rituals—such as meeting for coffee—consumers enact and reinforce social bonds, local identities, and communal belonging (Reckwitz, 2002). For example, in Italy, drinking coffee in cafés is a social practice that sustains community identity and embodies collective values of sociability and conviviality.

Consumption and identity theory

Ritual consumption serves as a means of constructing, expressing, and maintaining individual and collective identities. From this perspective, consumption rituals operate as symbolic resources through which individuals communicate social roles and group affiliations (Belk, 1988). Products used in ritual contexts become markers of lifestyle and social differentiation. The same product—coffee—may signify distinct identities depending on the context of consumption (e.g., at home, at work, or in a premium café). Consuming branded coffee such as Starbucks can signal cosmopolitan identity and affiliation with global consumer subcultures.

Symbolic consumption theory

Products possess symbolic meanings that transcend their functional utility. Symbolic consumption theory posits that consumption rituals operate as sign systems through which individuals convey meaning and status (McCracken, 1986). Ritual acts, such as preparing or serving coffee, communicate aesthetic and cultural values, and the choice of

specific products or preparation methods can express prestige, taste, or refinement. Brewing high-quality single-origin coffee using specialized equipment can symbolize cultural sophistication and connoisseurship.

Consumption and emotion theory

Ritual consumption elicits and regulates emotions, which, in turn, shape consumer behavior. This approach emphasizes the affective dimension of consumption, framing rituals as emotionally charged practices that provide comfort, pleasure, or psychological stability (Holbrook and Hirschman, 1982). Emotional gratification derived from ritualized acts—such as morning coffee routines—reinforces habitual engagement and contributes to subjective well-being. For many individuals, drinking coffee in the morning induces a sense of calmness and readiness, transforming a functional act into an emotionally meaningful ritual.

Consumption as cultural ritual theory

Ritual consumption is embedded in cultural traditions that shape collective understandings of products and their use. From a cultural anthropological perspective, ritual consumption transmits shared values and traditions across generations (Douglas and Isherwood, 2002). Coffee rituals, for instance, reflect broader cultural cosmologies and moral orders, dictating how coffee should be brewed, served, and consumed. Such practices both preserve heritage and adapt to modernity. Like in Ethiopia, the traditional coffee ceremony exemplifies cultural continuity, where the ritualized preparation and communal sharing of coffee express hospitality, respect, and social cohesion.

Research methods

Research process and sample structure

The objective of this research is to develop and validate the Coffee Ritual Consumption Scale. The study was conducted on a convenience sample of 128 university students from the Wyższa Szkoła Finansów i Bankowości (Higher School of Finance and Banking) in Kraków, Poland. The use of convenience sampling was determined by the methodological focus of the research, which aimed to (a) assess the reliability of the Coffee Ritual Consumption Scale, and (b) evaluate a new composite quality index integrating traditional reliability measures for individuals and items with indicators of outlier and inlier responses within the scale structure[1].

The initial version of the Coffee Ritual Consumption Scale consisted of 12 items reflecting ritualized patterns of coffee consumption. The original items included:

1. RCS 1 - Drinking coffee is part of my daily routine.
2. RCS 2 - I usually drink coffee at the same time each day.
3. RCS 3 - Preparing or going for coffee feels like a small ceremony that I celebrate.
4. RCS 4 - A coffee break is a moment during the day when I slow down and truly relax.
5. RCS 5 - Drinking coffee is one of the most important moments of my day.

[1] I would like to thank Professor Magdalena Kowalska-Musiał from WSFiB and Doctor Anna Myrda from University of Applied Sciences in Nowy Targ for their help in conducting empirical research.

6. RCS 6 - While drinking coffee, I savor its taste and aroma.
7. RCS 7 - I feel discomfort when I cannot drink coffee in my preferred way.
8. RCS 8 - I usually drink coffee prepared in the way I like most.
9. RCS 9 - I have my favorite place for drinking coffee.
10. RCS 10 - The morning cup of coffee signals that a new day has begun.
11. RCS 11 - The last cup of coffee in the evening closes my day and helps me unwind.
12. RCS 12 - I cannot imagine myself without my daily cup of coffee.

The 12 items presented above constituted the foundations for IRT-based scale reliability and quality analysis.

Person and item scale reliability and quality indices

One of the advantages of IRT scales is to provide more complex reliability and scale quality indices in comparison to the classical test theory (CTT) counter-part. Because of the cumulative nature of both items (item difficulty) and objects (person ability), the reliability coefficients are based on both person fit and item fit. Contrary to CTT, reliability and quality metrics are not constant, but vary across the persons' ability continuum. One can distinguish several dimensions and aspects of scale quality: 1/ global fit, 2/ local fit, 3/ scalability, 4/ reproductability, 5/ separation, 6/ information, 7/ reliability.

The scale quality indices are summarized in table 1.

Table 1. Scale quality indices

Dimension	Purpose	Indices	Focus
Local Fit	Detect item–person misfit	Infit, Outfit	Consistency of responses
Separation	Distinguish latent trait levels	Person Separation Index, Item Separation Index	Test sensitivity to variation
Reliability	Evaluate stability and true-score variance	Person / Item / EAP / WLE reliability	Precision of measurement
Information	Assess conditional measurement accuracy	Item information function Test Information Function	Precision at specific θ
Scalability	Assess cumulative ordering	Loevinger H (Hi, Hij), Construct Scalability	Ordinality and monotonicity
Reproducibility	Test Guttman pattern consistency	Construct Reproductability, Marginal MR	Degree of deterministic structure
Global Fit	Evaluate model adequacy	χ^2, RMSEA, SRMR	Overall measurement model fit

Source: own study.

These indices can be arranged into a hierarchical structure of fit:

1. Lower tiers (1–2) - can detect local anomalies in the data.
2. Middle tiers (3–4) quantify measurement precision and information of scale.
3. Upper tiers (5–7) evaluate overall structural and model-level measurement quality.

Together, they define a hierarchical quality system linking response behavior, measurement reliability, and the theoretical coherence of the scale. In empirical validation of scale, we propose a bottom-up approach from lower to upper tiers (local fit - reliability - scalability - global fit). In scale refinement, a top-down approach is suggested (theoretical expectations - empirical diagnostics). In mixed frameworks (e.g., non-parametric Mokken scaling and parametric IRT model), tiers 1–4 apply to latent measurement precision, while 5–6 address cumulative structure and scale monotonicity.

Local fit is measured by infit and outfit indices both for individuals and items within IRT framework of measurement quality.

1. Infit is the information-weighted measure of local fit between the observed and expected response patterns. It reflects how well an item (or person) fits the IRT model, giving greater weight to responses close to the person's estimated ability level (\). Person fit is given by the formula:

$$\text{Infit}_{\text{person } n} = \frac{\sum_i P_{ni}(1-P_{ni})(O_{ni}-P_{ni})^2}{\sum_i P_{ni}(1-P_{ni})} \quad (1)$$

Where:

O_{ni} - observed response of person *n* to item *i*

P_{ni} - expected response of person *n* to item *i* in the IRT model

Persons infit indicates how consistent the person's response pattern is for items near their own ability level (so for the inliers).

Item fit is given by the formula:

$$\text{Infit}_{\text{item } i} = \frac{\sum_n P_{ni}(1-P_{ni})(O_{ni}-P_{ni})^2}{\sum_n P_{ni}(1-P_{ni})} \quad (2)$$

Item infit indicates how consistently the item functions across respondents with abilities near its difficulty.

2. Outfit is a non-weighted mean-square residual that assesses the extent to which the observed responses deviate from those expected by the IRT model, regardless of how informative each response is. It is particularly sensitive to outliers, i.e., unexpected responses to items that are very easy or very hard for a person, or to individuals whose responses are highly inconsistent with the model expectation.

Person outfit is calculated by the formula:

$$\text{Outfit}_{\text{person } n} = \frac{\sum_i (O_{ni}-P_{ni})^2}{\sum_i P_{ni}(1-P_{ni})} \quad (3)$$

The person outfit statistic detects whether person *n* provides unexpected responses, particularly to very easy or very hard items relative to their θ. A large outfit indicates an erratic or guessing behavior and a very low outfit shows overly consistent or patterned responses. The infit statistic is sensitive to unexpected responses near a respondent's ability level. Values of infit close to 1 indicate a good fit (Linacre, 2002).

Item outfit is calculated by the formula:

$$\text{Outfit}_{\text{item } i} = \frac{\sum_n (O_{ni}-P_{ni})^2}{\sum_n P_{ni}(1-P_{ni})} \quad (4)$$

The item outfit detects whether item *i* received unexpected responses from individuals far from its estimated difficulty (i.e. outliers).

A large outfit value (> 1.5) shows the item elicits unexpected answers (e.g., mis-keyed, ambiguous wording) and a small outfit (< 0.5) reflects too predictable and too regular items (redundant with other items and overfit). Values close to 1 indicate a good fit. (Bond and Fox, 2013).

The separation Indices consist of the Person Separation Index (PSI) and Item Separation Index (ISI). There are two key Rasch-based indicators that quantify the measurement precision and discriminatory power of a scale. They describe how well the test distinguishes between (a) individuals with different ability levels and (b) items with different difficulty levels — that is, how separated the measures are relative to their standard errors. Both indices are analogous to signal-to-noise ratios. They express the proportion of true variation exists relative to measurement error.

3. The Person Separation Index (PSI) indicates how well the test can discriminate between individuals with different ability levels. It reflects the ratio of true variance in person measures to error variance. A higher PSI means the test distinguishes more strata (levels) of ability among respondents. PSI is given by the formula:

$$\text{PSI} = \sqrt{\frac{SD_P^2 - SE_P^2}{SE_P^2}} \tag{5}$$

Where:

SD_p^2 - Variance of person estimated abilities ()

SE_p^2 - Mean squared standard error of person abilities

A higher PSI (>2.0), indicates better differentiation among individuals

4. Item Separation Index (ISI) indicates how well the sample of respondents can distinguish between items of different difficulty levels. ISI is given by the formula:

$$\text{ISI} = \sqrt{\frac{SD_I^2 - SE_I^2}{SE_I^2}} \tag{6}$$

Where:

SD_I^2 - Variance of item difficulties (b)

SE_I^2 - Mean squared standard error of item difficulty estimates

A higher ISI (>3.0) means item difficulties are well spread and precisely estimated.

The reliability of IRT scale is measured by several person and item indices like person, item, EAP and WLE reliabilities.

5. Person reliability quantifies how consistently the instrument distinguishes between individuals with different ability levels on the latent trait (). It is analogous to Cronbach's α, but derived from the Rasch model or other IRT frameworks. It estimates the proportion of observed variance in person measures that reflects true variance, rather than a measurement error. The formula of person reliability is given as:

$$\text{Person Reliability} = \frac{SD_P^2 - SE_P^2}{SD_P^2} \tag{7}$$

Where:

SD_p^2 - variance of person estimated abilities ()

SE_p^2 - mean squared standard error of person abilities

Person reliability is directly related to PSI:

$$\text{Person Reliability} = \frac{\text{PSI}^2}{1 + \text{PSI}^2} \tag{8}$$

A higher person reliability (>0.7) indicates higher discrimination among individuals.

6. Item reliability quantifies how well the sample distinguishes among item difficulties and indicates how stable and reproducible the item hierarchy would be across different samples. Item reliability reflects the precision of item calibration (not the consistency of respondents). The formula of item reliability is given as:

$$\text{Item Reliability} = \frac{SD_I^2 - SE_I^2}{SD_I^2} \tag{9}$$

Where:

SD_I^2 - variance of item difficulties (b)

SE_I^2 - mean squared standard error of item difficulty estimates

Item reliability is directly related to ISI:

$$\text{Item Reliability} = \frac{\text{ISI}^2}{1 + \text{ISI}^2} \tag{10}$$

A higher item reliability (>0.8) indicates higher calibration stability among items.

7. Expected a posteriori reliability (EAP reliability) is also known as R_1c reliability. EAP is used in Bayesian IRT estimation (e.g., MCMC or marginal maximum likelihood) and reflects the proportion of the total variance in posterior ability estimates explained by the true variance, averaged over the population. It can be regarded as the Bayesian analogue of person reliability, computed using posterior means and variances of θ estimates.

$$\rho_{EAP} = \frac{Var(\hat{\theta}) - \overline{V(\theta|X)}}{Var(\hat{\theta})} \tag{11}$$

Where:

$Var(\hat{\theta})$ - variance of EAP estimates across respondents

$\overline{V(\theta \mid X)}$ - mean posterior variance across all respondents in the sample, conditional on that person's observed response pattern X (the average measurement error variance). It represents the average uncertainty of ability estimates across respondents.

8. WLE reliability (Warm, 1989) is used when person abilities (ʖ) are estimated via Warm's Likelihood Estimate (WLE), a bias-corrected version of the Maximum Likelihood Estimate (MLE). It measures the ratio of true variance in WLE person estimates to total observed variance, similar to person reliability, but computed from WLE-based errors. The formula of WLE reliability is given as:

$$\rho_{WLE} = \frac{Var(\hat{\theta}_{WLE}) - \overline{SE^2_{WLE}}}{Var(\hat{\theta}_{WLE})} \tag{12}$$

Where:

$Var(\hat{\theta}_{WLE})$ - Warm's likelihood estimate of person ability

$\overline{SE^2_{WLE}}$ - mean squared error across respondents

EAP and WLE > 0.7 indicate acceptable reliability. EAP and WLE reliabilities often yield similar results but EAP tends to be slightly higher because it incorporates prior information.

Information quantifies the extent to which an item or test provides information about a person's latent ability (θ). Higher information means more precision and lower measurement error. Lower information indicates less precision and higher uncertainty (Brzezińska, 2020).

9. The item information function $I_i(\theta)$ (IIF) describes how much statistical information item i provides about ability level θ. Item information is an inverse function of the standard error of measurement (SEM) and the steepness of the item characteristic curve in IRT model. The formula of IIF is as follows:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}$$

(13)

Where:

$P_i(\theta)$ - probability of a correct (or "1") response to item *i* at ability level \

$P_i'(\theta)$ - First derivative of P_i(\) with respect to \

10. The Test Information Function (TIF) is the sum of item information functions across all *k* items in the test:

$$I_T(\theta) = \sum_{i=1}^{k} I_i(\theta)$$

(14)

TIF is highest where the test provides the most accurate ability estimates and indicates on the ability scale where the test is most informative.

Scalability coefficient (H) is the core statistic of Mokken Scale Analysis (MSA), which is a nonparametric item response model based on the principles of Guttman scaling (Mokken, 2011). It quantifies how well items form a monotone, cumulative, and unidimensional scale, i.e., how consistently higher latent trait levels correspond to higher item scores. Scalability reflects the degree of cumulative ordering among items and individuals in a scale. In Guttman terms, a perfectly scalable set of items means that if a person endorses a difficult item, they also endorse all easier items. In Mokken scaling (a probabilistic generalization of Guttman scaling), small deviations (errors) are allowed, and scalability quantifies how close the data are to the ideal cumulative pattern. Therefore, scalability measures the strength of monotone relationships among items or between items and the latent trait. The scalability is measured by Loevinger Hij, Hi and H coefficients. H_{ij} was given in equation (1).

11. Item scalability H_i for *k* items is given as:

$$H_i = \frac{\sum_{j \neq i} H_{ij}}{(k-1)}$$

(15)

And measures how well item i follows the cumulative ordering relative to the rest of the scale.

12. Finally, H is the average of all pairwise H coefficients across items in the scale:

$$H = \frac{\sum_{i<j} H_{ij}}{\text{number of item pairs}}$$

(16)

It represents the overall strength of the cumulative structure for the entire set of items. H > 0.3 indicates an acceptable cumulative scale.

Reproducibility quantifies how well a respondent's pattern of item responses can be predicted from the ordering of items on the latent variable. Reproducibility measures how closely the observed data approximate this ideal cumulative pattern and is defined as the proportion of responses that can be correctly predicted based on the deterministic Guttman model.

12. Coefficient of Reproducibility (CR) quantifies the proportion of correctly predicted responses if each person's answers were predicted solely from their total score on the scale. CR is the earliest and simplest indicator of scale quality (Guttman, 1944). CR is given by the formula:

$$CR = 1 - \frac{E}{N \times k}$$

(17)

Where:

E- total number of Guttman errors (observed responses that violate the cumulative pattern)

N - number of respondents

k - number of items

Coefficient of reproducibility (CR) > 0.8 indicates acceptable reproducibility (traditional Guttman criterion).

Global model fit refers to the degree the chi-square (χ^2) statistic measures the weighted sum of squared residuals between observed and expected (model-based) response frequencies or probabilities (sum of standardized squared residuals across all item–person pairs).

13. Chi-square measures how far the observed responses deviate from what the IRT model predicts, weighted by the Bernoulli variance:

$$\chi^2 = \sum_{n=1}^{N}\sum_{i=1}^{k}\frac{(X_{ni}-\hat{P}_{ni})^2}{\hat{P}_{ni}(1-\hat{P}_{ni})} \quad (18)$$

Where:

X_{ni} - observed response (0 or 1) of person *n* to item *i*

$\hat{P}_{ni}$ - expected probability of the correct response in the fitted IRT model.

Insignificant (p>0.05) and low chi-square statistic indicate a good overall model fit.

14. The Root Mean Square Error of Approximation (RMSEA) in IRT expresses the average discrepancy per degree of freedom between observed and expected responses, scaled for sample size. It is derived directly from the IRT χ^2 statistic:

$$\text{RMSEA} = \sqrt{\frac{\max(\chi^2 - df, 0)}{df(N-1)}} \quad (19)$$

Where:

df - degrees of freedom (number of cells) – (number of estimated parameters)

N - sample size

The RMSEA then expresses this misfit as a standardized measure per degree of freedom, giving a population-level approximation error per degree of freedom. RMSEA < 0.08 indicates a reasonable fit.

15. Localized item-level $|^2$ and item-level RMSEA indices assess local misfit for individual items. They examine whether the observed response proportions for an item across ability levels match the model-predicted probabilities for that item. They are localized goodness-of-fit tests applied to each item rather than the entire test. Orlando–Thissen $|^2$ on a local level is given as:

$$S\text{-}\chi_i^2 = \sum_{g=1}^{G}\frac{(O_{ig}-n_g\hat{P}_{ig})^2}{n_g\hat{P}_{ig}(1-\hat{P}_{ig})} \quad (20)$$

Where:

G - number of ability groups

i - specific item

Scale Quality Index - a new proposition

All fit and scale quality indices show separate aspects of the quality of IRT models. We propose new formulas for assessing quality of scale, combining reliability and local fit measures. The proposed Item Quality Index (IQI) and Person Quality Index (PQP) that quantify the psychometric quality of an item by integrating three distinctive dimensions:

(1) measurement error (item's standard error of measurement SEM),

(2) local model fit (through standardized Infit and Outfit residuals), and

(3) robustness to outliers (through a damping parameter c). The index can be expressed conceptually as:

$$\text{Quality}_i = \text{Reliability Component} \times \text{Fit Weight}$$

The first component (reliability) corresponds to the reliability-like term of the item, quantifying how much of its variance is free from measurement error. The second component, (fit weight) acts as a Gaussian fit weight, reducing the effective reliability as misfit increases. Thus, Quality_i can be interpreted as a weighted

reliability measure adjusted for local misfit. This hybrid measure produces a continuous indicator ranging between 0 and 1, where higher values denote better overall item or person quality.

The overall Quality_i index with two multiplicative components is given as:

$$\text{Quality}_i = \underbrace{(1 - \text{SEM}_i^2)}_{\text{precision}} \times \underbrace{\exp\left(-\frac{z\text{-Outfit}_i^2 + z\text{-Infit}_i^2}{c}\right)}_{\text{model fit penalty}} \quad (21)$$

The term $(1-\text{SEM}_i^2)$ expresses the relative precision of the item estimate. Small SEM values (high precision) produce values close to 1, while large SEM reduce the index towards 0. The exponential term introduces a penalty for local misfit: it decreases rapidly as standardized Infit or Outfit residuals deviate from 0, following a Gaussian-type decay. The scaling constant *c* modulates the penalty's strength — smaller *c* increases sensitivity to misfit, whereas larger *c* produces smoother damping. Table 2 presents the role and calibration of damping parameter c.

Table 2. Interpretation of c parameter

c	Effect	Interpretation
Small (< 3)	Strong penalty	Very strict criterion, suitable for high-stakes testing
Medium (4 - 8)	Moderate penalty	Balanced sensitivity (recommended default)
Large (> 8)	Weak penalty	Tolerant of local deviations, suitable for exploratory work

Source: own study.

The index integrates three fundamental aspects of item quality: 1/ measurement precision (reliability) that reduced the index when SEM is large, 2/ local model fit (z-infit/z-outfit) that penalizes unexpected response patterns (inliers and outliers) and 3/ robustness parameter c that regulates sensitivity to extreme or outlying misfit values. Empirically, c can be optimized by maximizing the correlation of Quality_i with item information or with global model fit indices. Hence, the index represents a hybrid psychometric criterion, combining the informational value of the item with its empirical conformity to the model.

Potential applications of Quality_i include item selection and ranking in adaptive testing or item banking (prioritizing high-quality items), the visualization of item quality across the latent trait continuum (e.g., Quality–θ maps), item screening during calibration, for identifying low-precision or misfitting items and aggregate test quality indices, e.g., mean or information-weighted averages of Quality_i.

The index theoretically ranges from 0 (poor quality) to 1 (excellent quality). We propose the following interpretation thresholds:

Table 3. Interpretation of Quality_i Index

Quality_i	Interpretation
0.00-0.20	Very low quality: high measurement error or severe misfit
0.21-0.50	Moderate quality: tolerable precision, some misfit
0.51-0.80	Good quality: precise and well-fitting item
0.81-1.00	Excellent quality: stable, informative, and well-fitting item

Source: own study.

For example, a highly precise, well-fitting item with SEMi=0.10 z-fit=0 yields $\text{Quality}_i \approx 0.99$, but a poorly fitting, imprecise item with SEMi=0.6, z-Infit=3 and z-Outfit=2.5 yields $\text{Quality}_i \approx 0.12$.

Table 4 presents basic psychometric properties of Quality_i index:

Table 4. Psychometric properties of Quality_i Index

Property	Interpretation
Scale	Continuous between 0 and 1
Monotonicity	Decreases monotonically as SEM or infit/outfit increase
Robustness	Controlled by the damping parameter c
Interpretation	Integrates precision, reliability, and model consistency into a unified metric

Source: own study.

The (Person) Quality_i can be extended to an Item Quality_j Index, which combines item-reliability and item-fit and is based on following formula:

$$\text{ItemQuality}_j = \frac{I_j(\theta^*)}{\max_j I_j(\theta^*)} \cdot \exp\left(-\frac{z\text{-outfit}_j^2 + z\text{-infit}_j^2}{c}\right) \quad (22)$$

Where:

$I_j(\theta^*)$ – item information for θ (i.e. mean or weighted density θ),

$z\text{-infit}_j$ and $z\text{-outfit}_j$ - standardized item's infit and outfit

c – penalty term

Person Quality assess the precision of ability and person-fit whereas (small SEM and small infit/outfit) Item Quality stresses the height of item information and item fit.

To sum up, the Person and Item Quality Index offer a unified, interpretable framework for evaluating item performance by combining the precision, model fit, and robustness in a single metric. Unlike traditional diagnostics that treat measurement error and misfit separately, this index captures their joint impact on item reliability and interpretability. Conceptually, it represents a fit-weighted reliability function consistent with modern item banking and adaptive testing frameworks, where both accuracy and model conformity are critical for item selection. Practically, the index can be used to flag low-quality items, guide recalibration, or monitor the stability of item parameters across administrations. Its continuous scaling (0–1) facilitates graphical inspection and threshold-based decision rules in psychometric quality control.

Although the Scale Quality Index provides an interpretable measure of psychometric performance, several limitations should be acknowledged. First, the functional form assumes independence between measurement error and misfit, whereas these components may interact in practice, particularly for multidimensional or locally dependent items. Second, the weighting parameter c requires empirical calibration; its optimal value may vary across instruments and sample sizes. Third, the exponential penalty assumes symmetric treatment of over- and underfit, which may not always be desirable. Future research should examine alternative weighting schemes, Bayesian formulations incorporating posterior uncertainty, and generalizations to polytomous or multidimensional IRT models. Cross-validation studies could further assess the index's robustness, interpretability, and criterion validity across diverse testing contexts and populations.

Research findings

Scale reliability and separation

The double monotonicity analysis within the Mokken framework confirmed that eight of these items formed a unidimensional cumulative scale, satisfying the assumptions of monotone homogeneity and invariant item ordering. Mokken Scale Analysis (MSA) is a nonparametric item response theory (IRT) model designed to identify unidimensional, ordinal, and cumulative scales (Mokken, 1971). Unlike parametric IRT (e.g., Rasch models), MSA does not assume a specific functional form of the item characteristic curve — only that the probability of endorsing an item increases monotonically with the latent trait. A key step in MSA is determining whether a set of items can be regarded as unidimensional and scalable, i.e., whether all items measure the same latent construct and follow a consistent cumulative order. The search normal algorithm is part of the Automated Item Selection Procedure (AISP) implemented in software such as R *mokken* library. Its purpose is to partition the item pool into one or more Mokken scales — each representing a potential uni-dimensional subscale — by examining the scalability coefficients (Hij) between all pairs of items. This procedure is especially useful when new scales are being tested to obtain a coherent and unidimensional subset of items before the validation. The search normal procedure operates in three main steps:

Pairwise scalability assessment - each pair of items is evaluated using the Loevinger's scalability coefficient (Hij), which indicates the strength and monotonicity of their relationship.

$$H_{ij} = 1 - \frac{E_{ij}}{O_{ij}} \tag{23}$$

Where E_{ij} is the expected number of Guttman errors in the item's independence assumption and Oij is the observed number of Guttman errors. Values of $H_{ij} > 0.3$ are considered acceptable (Mokken, 1971).

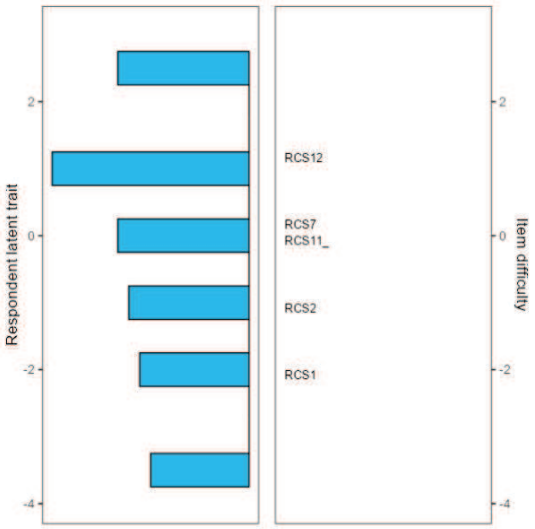
The algorithm is iterative, begins with the pair of items that have the highest pairwise scalability (Hij). It then iteratively adds items to this scale if their scalability with the existing items exceeds a minimum threshold (commonly c = 0.3). If an item does not meet this condition, it is not added to that scale, but may initiate a new one.

The algorithm continues until all items are assigned to one or more scales or excluded. The resulting set of scales reflects latent dimensions underlying the data. In the search normal variant, the algorithm performs the selection once, in a deterministic manner (as opposed to the search exhaustive method, which tests all possible combinations). Thus, "search normal" is a fast, rule-based method that yields a baseline dimensional solution, while "search exhaustive" is more computationally intensive but can find alternative or optimal item combinations. The algorithm continues until all items are assigned to one or more scales or excluded. The resulting set of scales reflects latent dimensions underlying the data. In the search normal variant, the algorithm performs the selection once, in a deterministic manner (as opposed to the search exhaustive method, which tests all possible combinations). The Hij coefficients with significance level () and dimensionality check [] for Mokken preliminary scale are as follows: RCS1: 0.351 (0.085) [1], RCS2: 0.269 (0.066) [1], RCS3: 0.294 (0.054) [2], RCS4: 0.261 (0.070) [2], RCS5: 0.351

(0.053) [1], RCS6: 0.168 (0.068) [0], RCS7: 0.316 (0.054) [2], RCS8: 0.312 (0.064) [2], RCS9: 0.135 (0.092) [0], RCS10: 0.290 (0.055) [1], RCS11: 0.416 (0.067) [1], RCS12: 0.311 (0.053) [1]. The scale is two-dimensional with some item not assigned to uni- or two-dimensional scale [0]. The final uni-dimensional scale consists of the following five items: RCS1: 0.731 (0.101) [1], RCS2: 0.513 (0.111) [1], RCS7: 0.549 (0.083) [1], RCS11: 0.533 (0.088) [1], RCS12: 0.464 (0.126) [1]. The scale H = 0.549 (0.077).

Figure 1 presents Wright map for both person abilities and item difficulties on the same latent trait scale (θ)

Figure 1. Wright map of RCS scale



Source: own study.

The figure exhibits a skewed distribution of latent trait. Items can be arranged along the difficulty scale from RCS1 (easiest item), RCS2, RCS11, RSC7 and RCS12 (most difficult item). The characteristic of scale items is that they cover -2 ; +2 standard deviation of scores with a lack of measurement of extreme abilities.

Table 5 presents the IRT reliability of coffee ritual consumption scale.

Table 5. Scale reliability

Measure	Reliability	Interpretation
Molenaar-Sijtsma	0.74	Acceptable reliability
Latent class reliability coefficient (LCRC)	0.82	Good reliability
Empirical reliability	0.73	Acceptable reliability
Marginal reliability	0.70	Acceptable re-liability

Source: own, based on *mokken* library of R package.

The overall reliability seems to be acceptable and ranges between 0.70 and 0.82.

Person and item separation indices are given in table 6.

Table 6. Person and item separability

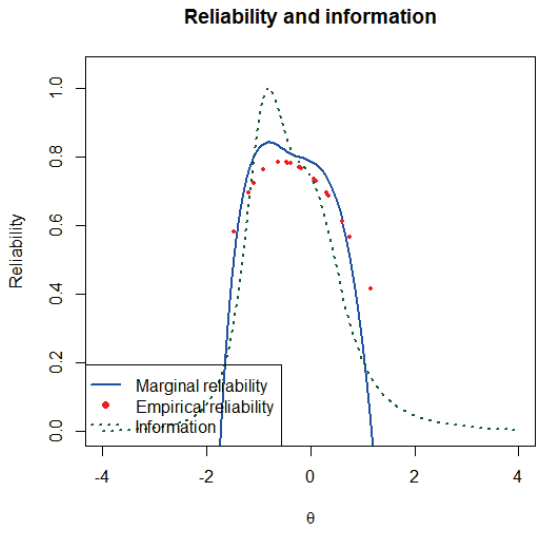
Measure	Index	Interpretation
Person Separation Index	0.99	Modest separation
Item Separation Index	NaN	Bad separation

Source: own study.

The scale separates respondents only modestly. Using the common rule of thumb, strata ≈ (4·PSI + 1)/3 ≈ 1.66, so barely 2 coarse groups can be distinguished at best (lower vs higher ritual consumption groups), with limited precision. The item separation index could not be estimated (NaN), indicating that the variance of item difficulties did not exceed their measurement error variance. This suggests that the five-item set did not provide a sufficient difficulty spread for reliable item calibration. Item calibrations are unstable (likely due to small N) and cannot be reliably ordered by difficulty. Their relative positions are uncertain.

Figure 1 depicts reliability and information along the ability latent trait (θ).

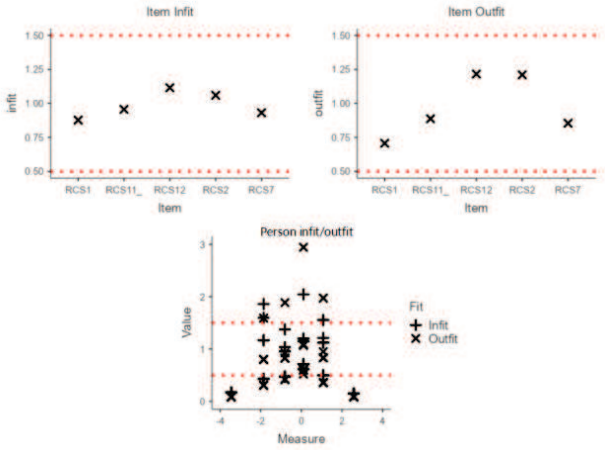
Figure 2. Reliability and information within latent trait range



Source: own study.

Figure 1 shows that the highest reliability, as well as information, is between -1 + 1.5 standard deviation from average score. It means that the scale performs especially well among less than average respondents and is most precise for respondents with moderate to slightly below-average levels of the latent trait. This means the scale is well-targeted toward people with typical ritual coffee behaviors. On the other hand, the scale lacks difficult items — it does not measure respondents with very high ritual engagement accurately. For those people, standard errors are large (low precision). The scale also lacks very easy items and low ritual engagement cannot be measured precisely as well. This result is consistent with the modest global person separation reliability (PSI ≈ 1.0). Specific person/item infit and person/item outfit plots are given in figure 3.

Figure 3. Infit and outfit indices



Source: own based on *jamovi* package.

Item-level fit statistics indicated good conformity with the Rasch model, with all infit and outfit mean square values falling within the acceptable range of 0.5–1.5. No item exhibited significant overfit or underfit, suggesting consistent item functioning. Person-level fit indices also demonstrated adequate model fit, with most respondents within the expected limits, and only a few outliers showing elevated outfit values (> 2.0). These results confirm the internal consistency and unidimensional structure of the ritual coffee consumption scale, supporting the appropriateness of the Rasch model for this item set. However, because the item separation index (ISI) could not be computed (NaN), this suggests that, although the model fits well, the items are not sufficiently dispersed in difficulty to provide distinct measurement strata. The test therefore functions reliably within a narrow ability range, but offers limited differentiation among higher- and lower-performing respondents.

Overall goodness of fit indices of Coffee Ritual Consumption Scale and comparison Rasch 1-parameter logistic (1PL) model and 2-parameter Birnbaum IRT model (2PL) are given in Table 7. It includes the Chi-square good-

ness of fit test, root mean square error of approximation (RMSEA), Akaike information criterion (AIC) and ANOVA difference test between 1PL and 2PL model

Table 7. Goodness of fit

Measure	Index	Interpretation
Chi-square 1PL	21.14, p = 0.68	1 PL Model match the data
RMSEA 1 PL	0	Perfect fit, very low root mean square error of approximation
AIC1 PL	407.98	Akaike information criterion for model comparison
Chi-square 2PL	15.24, p = 0.81	2 PL Model match the data
RMSEA 2PL	0	Perfect fit, very low root mean square error of approximation
AIC 2PL	410.08	Akaike information criterion for model comparison
ANOVA difference test 2PL - 1PL	Δ Chi-square = 5.89, p = 0.21	1 PL model is as good as 2PL model

Source: own study.

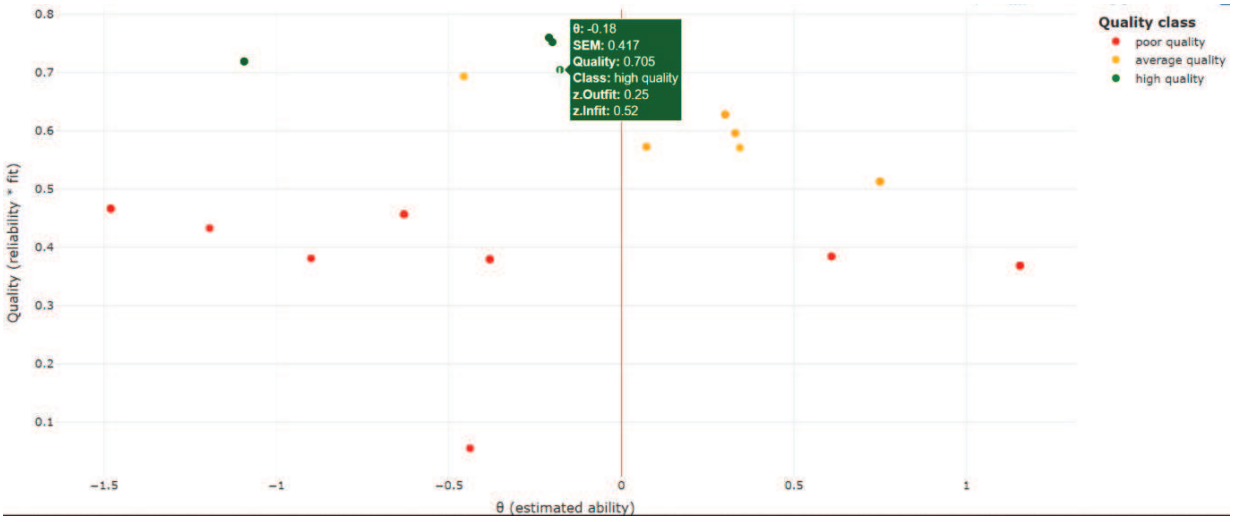
The overall goodness of fit is acceptable. Additionally, 1 PL Rasch model has an insignificantly worse fit in comparison to the more complex 2 PL Birnbaum model.

Person Quality$_i$ index

The result of Person quality index Quality$_i$ analysis across the latent trait is given in figure 4. For every respondent, the latent trait score (θ), his/her standard error of measurement (SEM), Quality$_i$, class of quality (class), standardized outfit (z-outfit) and standardized infit (z-infit) is calculated.

.

Figure 4. Person Quality$_i$ measures



Source: own study.

The $Quality_i$ index, representing fit-weighted reliability, demonstrated substantial variation across the latent continuum. Highest quality values (0.70–0.80) were observed around $\theta \approx -0.2$, where standard errors were minimal and model fit optimal, indicating precise and stable measurement in the midrange of ritual coffee consumption. Items and respondents at both extremes ($\theta < -1$ or $\theta > 0.5$) showed markedly lower quality (0.3–0.5), reflecting increased error and weaker fit. This pattern mirrors the test information function and confirms that the new $Quality_i$ index effectively integrates local reliability and fit. The analysis highlights the scale's strength in assessing average engagement levels, but its reduced precision for very low or high levels of ritual behavior.

Item Quality$_j$ index

A similar analysis of overall quality of scale concerning item quality using Item Quality Index $Quality_j$ is given in Table 8.

Table 8. Item quality index

Item	$Quality_j$	Interpretation
RCS1	0.991	High quality
RCS2	0.133	Low quality
RCS7	0.400	Average quality
RCS11	0.390	Average quality
RCS12	0.061	Low quality

Source: own study.

RCS1 demonstrates the highest item quality in the current scale configuration. Its relatively highest value (0.991) suggests that the item provides useful information for distinguishing between respondents and shows an adequate fit to the IRT model. It is likely a well-functioning item that contributes positively to the scale's psychometric properties. RCS2, with a Quality Index of 0.134, is markedly lower in quality. This indicates either low measurement precision (e.g., low information at relevant trait levels) or signs of misfit to the model. Further examination is warranted to assess its usefulness. RCS7 and RCS11 exhibit average item quality and RCS12 exhibit very low item quality, with values below 0.061. RCS2 and RCS12 appear to be poor contributors to the measurement model. The low index reflects a lack of sensitivity to the latent trait (e.g., items too easy or too difficult), poor discrimination and irregular response patterns (e.g., outlier behavior or guessing).

To further support scale refinement, future research should include inspecting the item characteristic and information curves, evaluating the scale's conditional reliability after excluding low-quality items and conducting a qualitative review regarding content validity of specific items.

Conclusions

The present study evaluated the psychometric properties of the Ritual Coffee Consumption Scale using item response theory. Results from the Rasch model indicated good model fit, but moderate overall measurement precision. All items demonstrated acceptable infit and outfit statistics (0.5–1.5), confirming local independence and unidimensionality. However, the Person Separation Index (PSI ≈ 1.0) and the inability to compute the Item Separation Index (ISI = NaN) suggested that, despite stable model calibration, the instrument provided limited differentiation among respondents and items. This outcome reflects a narrow clustering of

item difficulties and a relatively large standard errors of measurement. Consequently, although the model accurately reproduced observed response patterns, the current item configuration restricted the scale's capacity to discriminate among individuals with differing levels of ritual engagement.

The newly developed $Quality_i$ index, which integrates reliability and model fit, provided an additional insight into measurement performance. High-quality estimates ($Quality_i > 0.7$) were concentrated near the center of the latent continuum ($\theta \approx -0.5$ to 0.5), where standard errors were minimal and z-fit statistics were near expected values. In contrast, low-quality estimates appeared at the trait extremes ($\theta < -1.5$ or $\theta > 1.0$), reflecting increased uncertainty and weaker local fit. This pattern parallels the test information function, confirming that the scale measures moderate levels of ritual coffee consumption most effectively while providing limited information for respondents at very low or very high engagement levels. These findings support internal construct validity yet emphasize the need to broaden trait coverage for greater measurement sensitivity (Author, Year).

The scale's demonstrated model conformity and consistent reliability in the midrange of the trait suggest that it can serve as a valid tool for investigating everyday ritual practices within consumer and cultural domains. It may be applied to studies exploring lifestyle routines, symbolic consumption, and identity construction. However, reduced precision at the extremes indicates that interpretations of very low or highly ritualized behavior should be made cautiously. Expanding item coverage could improve discrimination, enabling researchers to segment consumer groups more effectively and enhance predictive models linking ritualized coffee consumption with well-being, social belonging, and cultural identity expression.

Several limitations must be acknowledged. The relatively modest sample size may have increased item standard errors, constraining item separation and precision at the tails of the latent continuum. The restricted range of item difficulties also limited the scale's capacity to represent the full spectrum of ritual engagement. Despite these constraints, the study exhibits key strengths: the application of rigorous item response modeling procedures, strong evidence of model fit and unidimensionality, and the introduction of the $Quality_{(i)}$ index as an innovative diagnostic that merges reliability and fit information. Collectively, these features provide a robust psychometric foundation for subsequent refinement and demonstrate the analytical utility of hybrid fit–reliability measures.

Future research should aim to extend the item pool to encompass both easier and more difficult statements, thereby achieving broader coverage across the latent trait continuum. Larger and more diverse samples would improve calibration stability and generalizability. Cross-cultural validation studies could examine the measurement invariance of ritual consumption behaviors across cultural contexts. In addition, the use of polytomous or multidimensional IRT frameworks may reveal distinct behavioral, emotional, and symbolic dimensions underlying ritual coffee consumption. Finally, longitudinal or experimental designs could explore how ritualized consumption develops over time and its potential contributions to psychological well-being, community integration, and cultural identity maintenance.

Declaration of competing interest

There are no financial and personal relationships with other people or organizations that could inappropriately influence (bias) our work.

References

Arnould, E. J., & Thompson, C. J. (2005). Consumer culture theory (CCT): Twenty years of research. *Journal of consumer research, 31*(4), 868-882.

Belk, R. W. (1988). Possessions and the extended self. *Journal of Consumer Research, 15*(2), 139-168.

Bond, T. G., Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences.* New York: Psychology Press.

Brzezińska, J. (2020). *Modele odpowiedzi na pozycje testowe (IRT) w badaniach społeczno-ekonomicznych*. Katowice: Wydawnictwo UE w Katowicach.

Constructive drinking: Perspectives on drink from anthropology. (1987). Douglas, M. (ed.), Cambridge University Press.

Douglas, M., Isherwood, B. (2002). *The world of goods: Towards an anthropology of consumption*. London: Routledge.

Guttman, L. (1944). A basis for scaling qualitative data. *American sociological review, 9*(2), 139-150.

Hattox, R. S. (2014). *Coffee and coffeehouses: The origins of a social beverage in the medieval Near East*. University of Washington Press.

Holbrook, M. B., Hirschman, E. C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of Consumer Research, 9*(2), 132-140.

Kucukkomurler, S., Ozgen, L. (2009). Coffee and Turkish coffee culture. *Pakistan Journal of Nutrition, 8*(10), 1693-1700.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Maciejewski, G., Mokrysz, S., & Wróblewski, Ł. (2020). *Consumers towards marketing strategies of coffee producers*. Wageningen: Wageningen Academic Publishers.

McCracken, G. (1986). Culture and consumption: A theoretical account of the structure and movement of the cultural meaning of consumer goods. *Journal of Consumer Research, 13*(1), 71-84.

Mokken, R. J. (2011). *A theory and procedure of scale analysis: with applications in political research*. Walter de Gruyter

Morris, J. (2013). Why espresso? Explaining changes in European coffee preferences from a production of culture perspective. *European Review of History, 20*(5), 881–901.

Oldenburg, R. (1999). *The great good place: Cafes, coffee shops, bookstores, bars, hair salons, and other hangouts at the heart of a community*. Da Capo Press.

Olejniczuk-Merta, A. (2024). The Impact of culture on coffee consumption: evolutionary perspective, *Coffee & Tea Marketing Journal, 1*(2), 25-38.

Ratcliffe, E., Baxter, W., L., Martin, N. (2019). Consumption rituals relating to food and drink: A review and research agenda, *Appetite, 134*(1), 86-93.

Reckwitz, A. (2002). Toward a theory of social practices: A development in culturalist theorizing. *European Journal of Social Theory, 5*(2), 243-263.

Rook, D. W. (1985). The ritual dimension of consumer behavior. *Journal of Consumer Research, 12*(3), 251-264.

Schatzki, T. R. (1996). *Social practices: A Wittgensteinian approach to human activity and the social*. Cambridge University Press.

Warde, A. (2005). Consumption and theories of practice. *Journal of Consumer Culture, 5*(2), 131–153.

Warde, A., Martens, L. (2000). *Eating out: Social differentiation, consumption and pleasure*. Cambridge University Press.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3).

Notes on contributors

Adam Sagan, professor of Management Sciences at Kraków University of Economics. Chair of Dept of Market Analysis and Marketing Research. Research interests: marketing research, latent variable modeling, means-chain chain analysis, IRT scaling.
email: sagana@uek.krakow.pl

Citation: Sagan, A. (2025). IRT Reliability of Coffee Ritual Behavior Scale - a New Metrics, *Coffee and Tea Marketing Journal*, *2*(2), 9-28. https: //doi.org/10.69102/CTMJ.2025.020201