

Dual Psychometric Evaluation of Primary School Teachers' Efficacy and Expectancy Beliefs in Mathematics Teaching

Abdellatif JHILAL

Sidi Mohamed Ben Abdellah University

Jaouad EL HILALY

elhilaly.jaouad@gmail.com

Sidi Mohamed Ben Abdellah University

Youness HADDER

Sidi Mohamed Ben Abdellah University

Khadija Saka

Sidi Mohamed Ben Abdellah University

Abderazzak MOUIHA

Euro-Mediterranean University of Fes

karim EL MOUTAOUAKIL

Sidi Mohamed Ben Abdellah University

Abdellatif BELCADI

Sidi Mohamed Ben Abdellah University

Adel TAHRAOUI

Sidi Mohamed Ben Abdellah University

Research Article

Keywords: Item Response Theory, Psychometrics, Validity, Outcome Expectations, Self-Efficacy, MTBI

Posted Date: June 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9557238/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Beliefs in mathematics teaching efficacy among future primary school teachers have been the subject of several studies. The Mathematics Teaching Efficacy Beliefs Instrument (MTEBI) is widely used in these studies. This study evaluates the structural validity and cross-cultural psychometric performance of the MTEBI within the Moroccan context.

Methods

The MTEBI was administered to 542 Moroccan elementary school student teachers. An exploratory factor analysis ($n = 105$) was conducted before a confirmatory factor analysis ($n = 542$) to test the two-dimensional structure of the instrument. An item response theory (IRT) model was then used to estimate the discrimination and threshold parameters of the items, as well as the information and conditional reliability curves.

Results

CFA corroborated the MTEBI's two-factor structure with acceptable fit. IRT analysis revealed that PMTE precision peaks for low-efficacy trainees ($q < 0$), whereas MTOE accuracy is localized near the mean ($\alpha_{\text{PMTE}} = .65$; $\alpha_{\text{MTOE}} = .80$). Item-level diagnostics identified threshold disorder for SE17, indicating response category misalignment.

Conclusion

The PMTE subscale identifies high-risk trainees, and MTOE reliability is localized near the mean. This mandates a transition from global CTT metrics to IRT-based conditional precision to ensure measurement equity in cross-cultural teacher education research.

1. Introduction

Teachers' self-efficacy theory conceptualizes instructional behavior as shaped by the dynamic interaction between teachers' beliefs in their capability to teach effectively (efficacy expectations) and their beliefs that such efforts will lead to desired learning outcomes (outcome expectations). Empirical evidence consistently indicates that the early stages of teacher preparation are critical for the development of strong self-efficacy beliefs (1, 2). Once these beliefs are established, teachers' self-efficacy during in-service practice appears to be comparatively stable and less amenable to change than that of preservice teachers (3).

Teacher efficacy is widely conceptualized as domain-specific, with educators' confidence differing across subject areas(4). This issue is particularly salient in the Moroccan educational context, where primary school teachers function as generalists responsible for educating learners aged 4 to 13. These teachers are required to deliver the full curriculum, often in the absence of specialized training or substantive pedagogical expertise in mathematics and mathematics instruction.

Mathematics-focused pedagogical coursework and practice-based teaching experiences are key determinants of preservice teachers' instructional efficacy (5, 6). Teaching efficacy is inversely associated with mathematics anxiety (7, 8), a relationship that has direct implications for instructional quality. Preservice teachers with lower efficacy tend to emphasize procedural memorization and repetitive exercises, often at the expense of problem-solving (9, 10). In contrast, those with stronger efficacy beliefs are more likely to enact cognitively demanding tasks, promote mathematical discourse, and foreground conceptual understanding (11)..

Mathematics teaching efficacy (MTE) crucially predicts primary teacher readiness and practice. While rooted in prior experience (12), MTE involves distinct personal and outcome-oriented beliefs(13) susceptible to training. Evidence shows MTE increases throughout university training (14), especially when bolstered by concrete practice, collaboration, and reflection. Thus, preservice efficacy beliefs are dynamic constructs that can be systematically improved through optimized learning environments.

Adapted from the Science Teaching Efficacy Beliefs Instrument (15), the MTEBI Instrument was validated by Enochs et al (16) using a multi-site sample of 324 preservice teachers. After the removal of two items with low item–total correlations ($r < 0.30$), confirmatory factor analysis (CFA) supported a stable two-factor structure comprising PMTE and Mathematics Teaching Outcome Expectancy (MTOE). Both subscales demonstrated strong internal consistency, with Cronbach's alpha coefficients of $\alpha = 0.88$ (PMTE) and $\alpha = 0.77$ (MTOE). The empirical independence of these dimensions provides further evidence of the instrument's construct validity for assessing mathematics teaching efficacy.

Nevertheless, cross-cultural evidence indicates that teaching efficacy beliefs vary markedly across national settings (17). This cultural sensitivity also extends to measurement validity. Although the MTEBI instrument has been adapted across diverse cultural contexts, including Arabic (18), Korean (19), Greek (20), Serbian (21), Spanish (22), and South African populations (23). These adaptations have not consistently reproduced the original two-factor structure. Such inconsistencies may reflect sample heterogeneity, particularly the aggregation of primary and secondary teacher cohorts, underscoring the need for level-specific investigations. Moreover, recent gender-focused analyses reveal significantly lower self-efficacy among male preservice-teachers (PSTs) (23), highlighting the need for targeted interventions that integrate rigorous pedagogy, sustained clinical experiences, and structured mentorship to address demographic disparities in teacher preparation.

Historically, validation of the MTEBI has been grounded in Classical Test Theory (CTT) (16). However, CTT is constrained by fundamental limitations, including uniform measurement precision across the

latent trait continuum and dependence of item parameters on the calibration sample. As a result, person scores and item characteristics are confounded, necessitating repeated revalidation across populations.

Item Response Theory (IRT) provides a more robust framework by modeling the probabilistic relationship between latent traits and item responses (24). By estimating item difficulty and discrimination on a common latent scale (24, 25), IRT produces sample-invariant parameters and trait-level reliability estimates, supporting more precise score interpretation and stronger construct validity. Moreover, IRT enables advanced psychometric applications unavailable under CTT, such as differential item functioning analyses, test equating, and computerized adaptive testing. Consequently, IRT has increasingly supplanted CTT as the preferred approach for evaluating measurement structure and validity in contemporary educational research (26, 27).

The evolution of IRT models reflects a shift from dichotomous performance metrics to complex polytomous structures. For Likert-type scales, the Graded Response Model (GRM) remains the standard for robust parameter estimation (28). GRM is uniquely suited for Likert-type data, as it accommodates items with varying response categories without compromising parameter estimation or interpretability. As a psychometric extension of the 2PL model (24), the Graded Response Model (GRM) (29) treats an item with m_i categories as a series of $m_i - 1$ cumulative dichotomous contrasts. For a 5-point Likert item, four progressive dichotomies are modeled (28).

Mathematics teacher training at the Regional Center for Education and Training Careers (CRMEF) is grounded in a competency-based framework that integrates theoretical knowledge with classroom practice. Central to the curriculum is mathematics didactics, drawing on Shulman's Pedagogical Content Knowledge (PCK) and the Mathematical Knowledge for Teaching (MKT) framework to support the transformation of complex mathematical concepts into teachable content. This foundation is complemented by coursework in educational sciences, including developmental psychology and sociology, to promote learner-centered instruction.

A distinctive feature of the program is its integration of digital pedagogy through the TPACK framework, preparing trainees to employ ICT tools such as GeoGebra and artificial intelligence in mathematics teaching. A professional qualification also requires an action-research project, fostering reflective practice on authentic pedagogical challenges. To support professional development, CRMEF monitors trainees' self-efficacy using standardized instruments such as the MTEBI.

This study examines the structural validity and psychometric properties of the MTEBI in Moroccan primary education, combining Classical Test Theory (CTT) and Item Response Theory (IRT). Specifically, it addresses the following research questions :

- Q1: Does the MTEBI demonstrate structural validity and reliability among Moroccan preservice primary teachers within the CTT framework?
- RQ2: Does the GRM model exhibit adequate model-data fit for the MTEBI among Moroccan preservice primary teachers?

- RQ3: To what extent do TIFs and item parameters establish the reliability and construct validity of the MTEBI within the Moroccan context?
- RQ4: How do IRT findings complement, qualify, or enrich CTT-based conclusions regarding the instrument's psychometric properties?

2. Methods

2.1 Research design

A cross-sectional design was conducted between March and June 2024 to investigate the structural validity and psychometric properties of the MTEBI in Moroccan primary education, combining Classical Test Theory (CTT) and Item Response Theory (IRT). The study was approved by the Committee the CRMEF administration

2.2 Participants and Setting

Participants were preservice elementary teachers enrolled at CRMEF Fez and Oujda. Data were collected at the end of a semester-long mathematics methodology course. The study employed a two-phase design: (1) exploratory factor analysis (EFA) on a pilot sample (N = 105) to establish the preliminary structure, and (2) confirmatory factor analysis (CFA) on an independent sample (N = 542) to validate dimensionality. Item Response Theory (IRT), using the Graded Response Model (GRM), was subsequently applied to examine item-level properties and model the latent dimensions of self-efficacy and outcome expectancy.

2.3 Instrumentation and Adaptation

The MTEBI, comprising two dimensions - Self-Efficacy (SE; 13 items) and Outcome Expectancy (OE; 8 items) - was adapted to the Moroccan context following cross-cultural validation guidelines. The English version was translated into Arabic by bilingual specialists, reviewed by an expert panel, and back-translated to ensure semantic and conceptual equivalence. The final Arabic version was pilot-tested with 20 preservice teachers, who reported no comprehension issues; average completion time was 15 minutes. No further modifications were required.

2.4 Procedure and Data Collection

A hybrid strategy (paper and Google Forms) was employed, featuring demographic items and the 21-item MTEBI core on a 5-point Likert scale. Ethical protocols were strictly followed: all participants provided informed consent after being briefed on the study's objectives, voluntary nature, and guaranteed anonymity and confidentiality.

2.5 Sample size estimation

The pilot sample (N = 105) satisfied the 5:1 participant-to-item ratio required for preliminary EFA (30). For the main phase, an independent sample (N = 542) was recruited; this exceeds the 500-participant

threshold recommended for stable CFA and polytomous IRT (GRM) modeling. After cleaning, the final analytic sample (N = 513) fully met international standards for stable factor solutions and reliable item parameter estimation(30, 31).

2.6 Data analysis

Data were cleaned to ensure accurate latent estimation, with items coded so that higher scores reflect greater teaching efficacy. To minimize bias, cases were excluded based on response inconsistency (straight-lining), boxplot-identified outliers [$Q1 - 1.5IQR$, $Q3 + 1.5IQR$], and aberrant patterns across $k \geq 4$ items. This process removed 29 cases (5.3%), yielding a final analytic sample of 513 preservice teachers (94.7% retention) for robust factor and IRT analyses.

2.7 Statistics

Analyses were conducted in R (v.4.5.2) using psych, lavaan, and mirt, following a dual CTT-IRT approach. EFA (N = 105) utilized Principal Axis Factoring (oblimin) and Parallel Analysis. CFA (N = 513) used Robust DWLS, with fit assessed via c^2 , Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA); Cronbach's Alpha (α), Composite Reliability (CR), Average Variance Extracted (AVE), and Heterotrait-Monotrait Ratio (HTMT) established validity and reliability(32, 33). The Graded Response Model (GRM) calibrated subscales independently after verifying unidimensionality and local independence (Yen's Q3). Precision was evaluated through Test Information Functions (TIF) and Category Response Curves (CRCs) across the latent continuum.

2.8 Ethical consideration

This study This study was approved by the Research Ethics Committee of the "Centre Régional des Métiers d'Éducation et de Formation, Fès Meknès" (approval code: CRMEF-30-02-2024, 02 /24). All procedures performed in this study were in accordance with the ethical standards of the CRMEF's Research Ethics Committee and with the 1964 Helsinki declaration and its later amendments. Informed consent was obtained, assuring participants of data confidentiality, of their right to withdraw anytime and of the fact that this

3. Results

Sample characteristics

The study included two independent samples of trainee primary school teachers: an exploratory sample (N = 105) used for EFA and a larger confirmatory sample (N = 542) used for CFA and IRT analyses. In accordance with standard validation practice, only the sociodemographic characteristics of the confirmatory sample were reported (Table 1). Most participants were recruited from CRMEF Fez–Meknes (78%), with the remainder from CRMEF Oujda (22%). The majority were aged 20–25 years (59.4%) and female (76%), with no significant interregional differences in age ($p = .10$) or sex distribution ($p = .25$). Most respondents held a bachelor's degree (89%), although regional differences were observed for degree level ($p = .047$) and professional bachelor's attainment ($p = .007$). Teaching experience was

generally limited, with 58% reporting no prior experience and 79% no experience teaching mathematics; no significant regional differences were found ($p > .05$). Overall, the absence of meaningful variation across most variables supports the comparability of subsamples and reinforces the internal validity of subsequent psychometric analyses.

Table 1
Sociodemographic characteristics of the participants

Gender	Total N = 542¹	Fez-Meknes N = 423¹	Oujda N = 119¹	p-value²
				0.25
<i>Female</i>	413 (76%)	327 (77%)	86 (72%)	
<i>Male</i>	129 (24%)	96 (23%)	33 (28%)	
Teachers' age				0.10
<i>20–25</i>	322.0 (59.4%)	259.0 (61.2%)	63.0 (52.9%)	
<i>26–30</i>	220.0 (40.6%)	164.0 (38.8%)	56.0 (47.1%)	
Marital status				0.64
<i>Single</i>	454 (84%)	356 (84%)	98 (82%)	
<i>Married</i>	88 (16%)	67 (16%)	21 (18%)	
Highest degree				0.047
<i>Bachelor</i>	483 (89%)	371 (88%)	112 (94%)	
<i>Master</i>	59 (11%)	52 (12%)	7 (5.9%)	
Professional Bachelor's degree				0.007
<i>No</i>	520 (96%)	401 (95%)	119 (100%)	
<i>Yes</i>	22 (4.1%)	22 (5.2%)	0 (0%)	
Professional experience				0.13
<i>1–4 year</i>	90 (17%)	77 (18%)	13 (11%)	
<i>< 1 year</i>	123 (23%)	93 (22%)	30 (25%)	
<i>> 4 year</i>	14 (2.6%)	13 (3.1%)	1 (0.8%)	
<i>Without</i>	315 (58%)	240 (57%)	75 (63%)	
Type of experience setting				0.50
<i>Private</i>	212 (39%)	171 (40%)	41 (34%)	
<i>Public</i>	22 (4.1%)	17 (4.0%)	5 (4.2%)	
<i>None</i>	308 (57%)	235 (56%)	73 (61%)	
² Pearson's Chi-squared test; Fisher's exact test				

Gender	Total	Fez-Meknes	Oujda	p-value²
	N = 542¹	N = 423¹	N = 119¹	
				0.25
Experience in teaching mathematics				0.15
No	430 (79%)	330 (78%)	100 (84%)	
Yes	112 (21%)	93 (22%)	19 (16%)	
¹ n (%)				
² Pearson's Chi-squared test; Fisher's exact test				

Exploratory factor analysis

The initial factor structure of the MTEBI was examined using a convenience sample of 105 trainees. The correlation matrix was adequate (KMO = 0.75; Bartlett's $\chi^2 = 375$, df = 91, $p < .001$) (30). Mardia's test indicated non-normality, so parallel analysis based on polychoric correlations was conducted, confirming a two-factor solution. Exploratory factor analysis (PAF, oblimin rotation) with iterative item purification yielded a stable structure (34,35): nine items loading on Self-Efficacy (eigenvalue = 3.2) and five on Outcome Expectancy (eigenvalue = 1.2), explaining 45% of variance. Reliability was modest for Self-Efficacy ($\alpha = .65$) and good for Outcome Expectancy ($\alpha = .80$), with total scale $\alpha = .67$, acceptable for exploratory research (Table 2).

Table 2
Factor loadings for Exploratory Factor Analysis with oblique rotation.

Items	Factors		Alpha	
	PMTE	MTOE	Inter-items	Total
SE3	0.66	0.06	0.65	0.67
SE5	-0.72	-0.15		
SE6	0.43	-0.31		
SE8	0.5	-0.39		
SE11	-0.76	-0.08		
SE15	0.34	0.1		
SE16	-0.56	0.17		
SE17	0.69	-0.13		
SE20	-0.44	-0.02		
OE1	0.15	0.33	0.80	
OE4	0.04	0.37		
OE9	0.04	0.37		
OE12	-0.01	0.48		
OE13	0.02	0.5		
Valeur propre	3.2	1.2		
Variance totale (%)	45			

Confirmatory factor analysis

Convergent validity

The results of the first-order CFA show that the majority of standardized factor loadings exceeded 0.50 and were statistically significant ($p < .001$), thus providing initial empirical evidence of convergent validity (30).

The hypothetical bifactorial structure of the 14-item MTEBI (PMTE and MTOE) was broadly confirmed. However, examination of CR and AVE indices suggests moderate convergent validity. Although CR is considered a more robust and less biased estimate than Cronbach's alpha, the observed values (0.50 for PMTE and 0.60 for MTOE) remain below the recommended threshold of 0.70, indicating limited internal consistency. In contrast, the MEV values (0.52 for PMTE and 0.54 for MTOE) exceed the minimum

threshold of 0.50, suggesting that, despite moderate composite reliability, the constructs explain more than 50% of the average variance of their respective indicators. These results suggest acceptable but improvable convergent validity (Table 3).

Table 3
Composite reliability (CR), average extracted variance (AVE), and Heterotrait-Monotrait ratio of correlations (HTMT).

Latent Constructs	CR	AVE	$\sqrt{\frac{\text{varvec{A}}}{\text{varvec{V}} + \text{varvec{E}}}}$	Correlations		HTMT	
				1	2	1	2
1.PMTE	0.5	0.52	0.71	0.71		1	
2.MTOE	0.6	0.54	0.73	0.3**	0.73	0.32	1

PMTE: Personal Mathematics Teaching Efficacy, MTOE: Mathematics Teaching Outcome Expectancy.

Discriminant validity

Discriminant validity was examined using two complementary approaches: the criterion developed by Fornell and Larcker (36) and the HTMT (ratio). The diagonal values in the table correspond to the square root of the AVE for each construct (0.71 for PMTE and 0.73 for MTOE). These values are higher than the observed interfactor correlation ($r = 0.30$), which satisfies Fornell and Larcker's criterion and indicates that each construct shares more variance with its own indicators than with the other construct. Furthermore, the HTMT value obtained (0.32) was below the recommended threshold of 0.85 (31), confirming the existence of adequate discriminant validity between the two dimensions of the model (Table 3). Taken together, these results indicate that the dimensions of the MTEBI measure related but distinct concepts.

Fitness of the measurement model

The evaluation of the first-order measurement model was based on several categories of fit indices: absolute indices (RMSEA), incremental indices (CFI, TLI), and parsimony indices (χ^2/df). The statistics obtained are as follows: $\chi^2(76) = 176$, $p = 0.06$; $\chi^2/\text{df} = 2.3$; RMSEA = 0.05; TLI = 0.90; CFI = 0.89 (Table 4). The χ^2/df ratio of less than 3 indicates an acceptable fit. The RMSEA of 0.05 suggests a good absolute fit of the model. The incremental indices (CFI = 0.89; TLI = 0.90) are slightly below the optimal threshold of 0.95 (37), but remain close to the minimum threshold of 0.90 generally accepted in the literature. Overall, the indices fall within acceptable ranges, suggesting that the model fits the data moderately well. The measurement model, therefore, presents a coherent structure that is generally well-adjusted for the MTEBI scale (Table 4).

Table 4
Confirmatory Factor Analysis Model Fit Indices

Fit Indices	Obtained Value	Recommended Threshold
Chi-square (df)	176 (76)	< 5
p-value of the Chi-square test	0.06	> 0.05
Root Mean Square Error of Approximation (RMSEA)	0.05	< 0.08
Tucker-Lewis Index (TLI)	0.90	> 0.9
Comparative Fit Index (CFI)	0.89	> 0.9

Checking the IRT assumptions

The psychometric properties of the MTEBI scale were examined using the GRM model within the framework of IRT. The structural validity of the instrument was rigorously assessed using a sequential factorial approach. The IRT assumptions regarding the dimensionality of the constructs were confirmed through EFA and CFA, while local independence was assessed using Yen's (1984) Q3 statistic. For the nine-item Self-Efficacy (SE) subscale, the expected Q3 value for independent item pairs was -0.125. Although most pairs adhered to the 0.30 threshold, significant deviations were observed for pairs SE5–SE8 (Q3 = -0.22) and SE11–SE15 (Q3 = -0.23), suggesting potential local item dependence. Conversely, the Outcome Expectancy (OE) subscale displayed evidence of negative local dependence in a single item pair (Tables 5 and 6).

Table 5
Q3 statistics for the personal efficacy subscale based on GRM parameter estimates

SE3	SE3	SE5	SE6	SE8	SE11	SE15	SE16	SE17	SE20
	1.000	-0.137	0.035	0.078	-0.178	-0.008	-0.246	0.013	-0.126
SE5		1.000	-0.164	-0.220	0.151	-0.137	-0.062	-0.131	-0.100
SE6			1.000	-0.002	-0.211	-0.006	-0.152	-0.083	-0.069
SE8				1.000	-0.152	-0.065	-0.206	-0.097	-0.094
SE11					1.000	-0.226	-0.013	-0.180	-0.113
SE15						1.000	-0.045	0.021	0.017
SE16							1.000	-0.113	0.031
SE17								1.000	-0.071
SE20									1.000
Note: The expected value for the Q3 statistics is -0.125. GRM = Graded Response Model.									

Table 6

Q3 statistics for the outcome expectations subscale based on MGR parameter estimates

OE1	OE1	OE4	OE9	OE12	OE13
	1.000	0.059	-0.049	-0.098	-0.165
OE4		1.000	-0.047	-0.217	-0.118
OE9			1.000	-0.072	-0.165
OE12				1.000	-0.412 ^a
OE13					1.000
Note: The expected value for the Q3 statistics is -0.25. GRM = Graded Response Model.					

a. Statistics greater than or equal to 0.30 in absolute value suggest local dependence.

Consistent with Ferrando et al. (38), local dependence in psychometric instruments often arises from semantic redundancy among items. In the MTEBI, item wording and spatial proximity appear to be key contributors to this effect. Within the Outcome Expectancy subscale, items OE12 and OE13 exhibit substantial conceptual overlap, both addressing student achievement in relation to instructional influence, and are positioned adjacently. OE12 emphasizes general teacher responsibility for student success, whereas OE13 focuses on the causal impact of instructional effectiveness. Their proximity may hinder respondents' differentiation of these nuances, resulting in correlated residuals.

A similar pattern was observed between Personal Efficacy items SE5 and SE8, likely driven by shared lexical content ("effective") and adjacency, despite opposite valences (positive vs. reverse-coded). As these dependencies were limited to a small number of item pairs, the assumption of local independence was not substantively violated. Nonetheless, the findings highlight the need for improved item differentiation to strengthen the instrument's structural validity.

Parameter estimation

An analysis of response option usage before parameter estimation revealed that, across the 14 retained items on the two subscales, "Strongly Disagree" (Response Option 1) was selected in no more than 3% of cases for twelve items on the MTEBI scale (Table 7). This pattern suggested a ceiling effect: participants appeared to perceive these items as very easy, favoring Category 2 or higher. Because sparse category frequencies can distort item parameter estimates (24), Categories 1 and 2 were collapsed for items SE3, OE4, SE5, SE6, SE11, OE12, OE13, SE15, and SE16, yielding four response options. Similarly, for items SE8 and SE20, the underutilization of Categories 1 and 2 extended to Category 3, prompting the consolidation of Categories 1–3 into a single option (39).

Table 7
Percentage of responses by category for each item

OE1	Categ :1	Categ :2	Categ :3	Categ :4	Categ :5
	4.68	25.15	20.86	37.04	12.28
SE3	1.95	7.60	10.33	42.11	38.01
OE4	0.78	5.07	5.85	53.41	34.89
SE5	0.39	9.36	20.27	50.88	19.10
SE6	0.39	4.68	14.81	56.73	23.39
SE8	0.58	0.78	5.07	38.40	55.17
OE9	6.04	7.41	9.16	45.03	32.36
SE11	1.36	7.21	8.97	42.88	39.57
OE12	1.95	12.48	18.52	48.54	18.52
OE13	1.17	20.08	19.30	47.17	12.28
SE15	1.56	14.04	17.93	50.68	15.79
SE16	0.97	4.29	15.40	59.65	19.69
SE17	4.09	13.65	10.33	37.82	34.11
SE20	0.97	0.97	2.34	32.55	63.16

The item parameters and associated standard errors for both subscales, derived via the Graded Response Model (GRM), are detailed in Tables 8 and 9. Following recent psychometric literature recommendations where slope parameters (a) exceeding 1.0 are categorized as acceptable (40), 64% of the instrument's items demonstrated sufficient utility in discriminating between varying levels of mathematics teaching efficacy. Discrimination parameters within the PE subscale ranged from 0.68 to 1.41. Applying De Ayala's classification (24), items SE11 and SE16 exhibited "high" discriminative power, while items SE3, SE5, SE6, SE8, and SE17 maintained moderate to high slopes ($1.03 \leq a \leq 1.29$). Conversely, items SE15 (0.68) and SE20 (0.93) displayed attenuated discrimination. This diminished performance is likely attributable to the necessary collapsing of response categories following low frequency counts, a process known to induce information loss and subsequent reduction in slope parameters (39). Despite these outliers, the PE subscale generally maintains a robust capacity to differentiate respondent efficacy levels.

The OE subscale exhibited a broader range of discrimination ($0.59 \leq a \leq 1.95$). Notably, items OE12 (1.85) and OE13 (1.95) demonstrated "very high" discrimination, serving as pivotal indicators for identifying subtle variances in the latent trait among high-performing respondents. These items represent the subscale's highest informational yield. In contrast, OE4 (0.79) provided only moderate

differentiation, while items OE1 (0.59) and OE9 (0.55) showed limited psychometric utility. In summary, while the OE subscale features high-performing "anchor" items, the presence of low-discrimination parameters suggests a degree of heterogeneity in how effectively individual items capture the underlying construct.

Analysis of the Personal Efficacy subscale revealed a robust distribution of item difficulty. The threshold parameters (b_i) spanned a broad range of the latent continuum, effectively capturing diverse levels of the underlying trait. Specifically, items SE5 and SE15 demonstrated significant variation in difficulty across ability levels. Conversely, item SE3 exhibited relatively low difficulty thresholds, indicating a high probability of endorsing higher response categories even among individuals with low latent efficacy.

Regarding the Outcome Expectancy subscale, difficulty thresholds ranged from -5.39 to 3.54 , with item OE1 encompassing both the minimum and maximum values. While discrimination parameters (a) dictate the peak information an item provides, the observed breadth of difficulty values is critical for maintaining measurement precision across all levels of θ . Given that IRT estimates are theoretically sample-independent, this distribution suggests a structurally sound instrument rather than a reflection of weak efficacy beliefs within the specific study cohort (Table 9).

Prototypical response patterns were modeled for a teacher exhibiting mean efficacy beliefs ($\theta = 0$). Such a teacher is likely to select the "agree" or "strongly agree" response categories for most items. The threshold parameters indicate that several items have relatively low difficulty levels, meaning that even respondents with an average level of self-efficacy are inclined to choose the positive response options (Tables 8 & 9).

Table 8
Item parameter estimates for the personal efficacy subscale

Item	a	SE	b_1	SE	b_2	SE	b_3	SE	b_4	SE
SE3	1.04	0.16	-2.54	0.15	-1.62	0.13	0.58	0.11		
SE5	1.09	0.16	-2.43	0.22	-0.96	0.13	1.61	0.16		
SE6	1.12	0.15	-3.08	0.22	-1.54	0.16	1.30	0.13		
SE8	1.29	0.26	-2.58	0.34	-0.23	0.14				
SE11	1.34	0.21	-2.25	0.27	-1.51	0.18	0.43	0.13		
SE15	0.68	0.11	-2.73	0.14	-1.15	0.10	2.67	0.13		
SE16	1.41	0.19	-2.60	0.27	-1.29	0.17	1.33	0.18		
SE17	1.03	0.14	-3.52	0.28	-1.87	0.15	-1.22	0.14	0.71	0.11
SE20	0.93	0.14	-3.72	0.25	-0.69	0.11				

Table 9
Estimates of item parameters for the outcome expectations subscale

Item	a	SE	b ₁	SE	b ₂	SE	b ₃	SE	b ₄	SE
OE1	0.59	0.11	-5.39	0.22	-1.60	0.11	0.01	0.09	3.54	0.14
OE4	0.79	0.13	-3.87	0.20	-2.85	0.16	0.90	0.11		
OE9	0.55	0.11	-5.22	0.19	-3.57	0.14	-2.39	0.11	1.42	0.11
OE12	1.85	0.28	-1.44	0.27	-0.60	0.17	1.22	0.24		
OE13	1.95	0.30	-1.07	0.23	-0.35	0.15	1.55	0.32		

Following parameter estimation, CRCs were analyzed for each item to characterize the relationship between the probability of selecting a specific response category and the latent self-efficacy continuum (41). As illustrated in Fig. 1 (Item SE16, about confidence in addressing students' mathematical inquiries), the curves depict distinct shifts in response probability relative to the latent trait (theta).

For instance, respondents with low personal efficacy ($q = -2.6$) exhibited an approximately 40% probability of selecting category 1 ('Strongly Disagree'), with a comparable likelihood for category 2 ('Uncertain'). In contrast, those at the higher end of the continuum ($q = +2.0$) showed a dominant probability (approximately 80%) of endorsing category 4 ('Strongly Agree'), with a near-zero probability of selecting the lowest categories.

The intersections of these CRCs signify the location of threshold parameters relative to personal efficacy levels. Furthermore, the spatial distribution of CRCs along the q continuum provides a diagnostic indicator of item difficulty. Specifically, items where CRCs cluster below the mean ($q < 0$) are characterized as relatively 'easy,' as they elicit high-category endorsements even from individuals with low latent efficacy.

Model Fit

We evaluated the fit of the Graded Response Model (GRM) at both the item and subscale levels. Item-level fit was quantified using Orlando and Thissen's $S-X^2$ statistic, where non-significant values indicate adequate model-data congruence (42, 43). As detailed in Table 10, all items exhibited non-significant $S-X^2$ statistics ($p > .05$), confirming that the GRM effectively characterized the item response patterns. While the global fit was robust, marginal performance in items SE15 and SE17 suggests localized areas for potential refinement. These findings validate the instrument's structural integrity and ensure that the latent variable estimates are derived from items with high psychometric fidelity.

The overall fit of the Graded Response Model to each subscale was evaluated. The results suggest that the model demonstrates an acceptable fit to the Personal Efficacy subscale and an excellent fit to the Outcome Expectancy subscale ($\chi^2(27, n = 513) = 116.59, p < .01$; RMSEA = 0.08, 90% CI [0.066, 0.096]; SRMR = 0.058; CFI = 0.831; TLI = 0.775; GFI = 0.946; AGFI = 0.91) and ($\chi^2(5, n = 513) = 8.053, p = 0.153$;

RMSEA = 0.035, 90% CI [0.000, 0.076]; SRMR = 0.026; CFI = 0.984; TLI = 0.968; GFI = 0.994; AGFI = 0.982), respectively.

Table 10
Item Fit Statistics (S-X²)

Item	S_X ²	df_s_x ²	p_s_x ²
SE3	39.417	32	0.172
SE5	18.173	31	0.967
SE6	37.330	29	0.138
SE8	10.740	20	0.953
SE11	20.785	28	0.834
SE15	56.010	34	0.010
SE16	21.358	26	0.723
SE17	77.904	38	0.000
SE20	11.034	20	0.945
OE1	31.965	28	0.276
OE4	19.564	23	0.668
OE9	40.231	30	0.100
OE12	23.595	21	0.313
OE13	21.298	19	0.321

Evidence of reliability and validity based on IRT

Measurement precision was evaluated through Item Information Functions (IIFs) and the Test Information Function (TIF), which provide a local assessment of reliability across the latent trait continuum (θ) (24). Under polytomous IRT frameworks, item-level information is the aggregate of category-specific information functions (44); consequently, items with higher discrimination parameters (α) yielded greater peak information and reduced standard errors of measurement (28).

Visual inspection of the IIFs (Figs. 2 and 3) facilitated the identification of the specific trait intervals where the scale achieves maximum psychometric sensitivity. While the *personal efficacy* dimension exhibited consistently moderate-to-high information density, indicating robust precision across the mid-range of θ , the *outcome expectancy* subscale demonstrated more heterogeneous information distributions. This variance in the latter subscale suggests that while certain items contribute

significantly to high-trait differentiation, others provide marginal utility, potentially narrowing the range of optimal measurement for this dimension.

A primary advantage of the IRT framework is the estimation of conditional reliability across the latent trait continuum. Figure 4 shows that the reliability curves for the MTEBI subscales reflect their respective information functions, thereby providing a localized assessment of measurement precision. The PE subscale demonstrates greater stability and measurement range than the OE subscale, maintaining robust reliability across a broader range of the latent trait (from $\theta = -3$ to $\theta = +2$). In contrast, the OE subscale achieves its maximum precision only near the mean, with its reliability gradually decreasing at the extremes ($\theta = -1$ à $+1$). These differences in reliability profiles constitute evidence of internal structural validity: they indicate that the two subscales do not function identically, which is consistent with their theoretical distinction. Furthermore, the convergence between scores estimated by the TRI and those derived from the TCT, regarding the relationship between the two subscales, provides additional support for the validity of the scores regardless of the scoring method used.

4. Discussion

Beliefs about teaching efficacy are central to effective mathematics instruction, yet their measurement remains contested due to concerns of validity and reliability. This study examined the psychometric properties of the MTEBI scale among Moroccan pre-service elementary teachers, applying both CTT and IRT to assess reliability, validity, and item performance across two samples in a distinct sociocultural context.

Exploratory Factor Analysis (EFA) necessitated the removal of seven items due to poor factor loadings, negative phrasing bias, and linguistic or cultural misalignments. Specifically, translation issues and differing cultural perceptions of teacher roles compromised the validity of these items. Their exclusion yielded a robust Arabic version that retains the original MTEBI's two-factor structure, personal efficacy, and performance expectancy, consistent with the framework established by Enochs et al. [55]. This structure has been repeatedly validated in previous studies (19–21, 45).

CFA of the refined instrument demonstrated satisfactory structural validity. While the CFI (0.89) was marginally below the conventional 0.90 threshold, the non-significant χ^2 test supported robust absolute fit. This profile aligns with established MTEBI validations (16, 21, 45) and reflects current psychometric shifts toward the contextualized interpretation of fit indices over rigid cutoffs (46). Consequently, these results confirm the structural integrity and cultural validity of the adapted Arabic version.

Construct reliability and validity assessments (CR and AVE) confirm the psychometric robustness of the Arabic MTEBI. Discriminant validity establishes that personal efficacy (PMTE) and outcome expectancy (MTOE) are empirically distinct, preserving the instrument's original conceptual structure. This dissociation underscores a critical nuance: trainees may perceive high personal competence independent of their beliefs regarding the impact of teaching on student outcomes. This documented

separation reinforces the necessity of treating these dimensions as complementary yet distinct constructs within teacher education frameworks (47, 48).

Following CTT confirmation, unidimensional Graded Response Models (GRMs) were applied to the PMTE and MTOE subscales. This bifurcation aligns with the Moroccan teacher training framework, which explicitly distinguishes between pedagogical self-efficacy and beliefs regarding teaching impact. Although local independence was generally satisfied, residual dependency between items OE12 and OE13 necessitates item rephrasing or increased spatial separation in future iterations to optimize item discrimination and reduce redundant variance.

Several items required collapsing response categories. Nine items necessitated grouping categories 1 and 2 (or 1–2–3 for SE8 and SE20), indicating that Moroccan trainees underutilized the lower extremes of the MTEBI. This pattern reflects both social desirability bias (49), with respondents reluctant to report low efficacy, and a cultural tendency to avoid negative self-judgments (50). Grouping categories is a recognized solution (24, 51) and recent studies show it improves model fit for sparse data without substantially reducing item information [49]. Nonetheless, it decreases scale granularity and highlights the need for a more thorough adaptation of verbal anchors to the Moroccan context, as recommended in other non-Western adaptations of the MTEBI (52, 53).

The graded response model (GRM) demonstrated superior fit for MTOE compared to PMTE. While MTOE showed excellent indices, PMTE exhibited only marginal fit, with notable misfit in items SE15 and SE17. This divergence likely stems from MTOE's conceptual homogeneity versus the inherent multidimensionality of self-efficacy (e.g., classroom management and student motivation). Within the Moroccan context, the emphasis on teaching impact over personal efficacy, coupled with the trainees' diverse backgrounds, further accounts for the increased variance in PMTE. These results corroborate that novice teachers' efficacy beliefs are highly context-sensitive (48), suggesting that while the GRM is effective for MTOE, the PMTE subscale requires further structural refinement.

The analysis of item parameters provides further evidence of the graded response model's performance. For MTOE, OE12, and OE13 displayed high discrimination ($a = 1.85$ and 1.95), effectively differentiating trainees with varying levels of outcome expectations. Their semantic proximity, however, explains the observed violation of local independence ($Q3 = -0.412$) (54). In contrast, discriminations for PMTE were moderate ($a = 0.68-1.41$). This indicates that trainees did not clearly distinguish between categories 2 and 3. The anomaly likely stems from translation choices, where direct and categorical wording reduced sensitivity to moderate responses. A more interrogative phrasing would better capture nuanced perceptions in the Moroccan context.

the category response curves for item SE16 (Fig. 1) illustrate one of the most discriminant items on the self-efficacy subscale ($a = 1.41$), further demonstrating the effective functioning of the graded response model. The curves activate sequentially, thresholds are correctly ordered ($b1 = -2.60$; $b2 = -1.29$; $b3 = 1.33$), and the highest category (P4) peaks at high levels of the latent trait ($\theta \approx 3$). This profile contrasts

with SE17, whose thresholds overlapped, and confirms that several MTEBI items function appropriately among Moroccan trainees.

IRT analyses confirm the MTEBI's psychometric robustness and conceptual validity within the Moroccan context. Test Information Functions (TIFs) and conditional reliability curves indicate that while both subscales are satisfactory, their precision profiles differ: the PMTE subscale offers broad measurement sensitivity—particularly for trainees with lower efficacy—whereas the MTOE subscale peaks near the mean and loses reliability at the extremes. These findings support the MTEBI's dual-factor structure while suggesting caution when interpreting extreme outcome expectancy scores. Specifically, the PMTE's high precision at lower latent trait levels is highly relevant for the Moroccan context, enabling the early identification of trainees requiring pedagogical support. Overall, IRT evidence validates the MTEBI as a reliable diagnostic tool, provided subscale-specific precision limits are acknowledged.

Comparative analysis of CTT and IRT frameworks reveals significant psychometric synergy. While both models confirm the MTEBI's two-dimensional structure and general item validity, a critical divergence in reliability emerged. CTT indicates superior reliability for the Outcome Expectancy (OE) subscale ($\alpha = 0.80$); however, IRT demonstrates that this precision is localized near the mean, whereas Personal Efficacy (PE) maintains higher accuracy across a broader latent trait range (θ). Furthermore, IRT provides essential diagnostic granularity by identifying specific item anomalies (SE17, OE1, OE9) and cross-cultural misalignments obscured by CTT's global indices. Ultimately, CTT offers a robust normative overview, while IRT facilitates the nuanced refinement required for adapting the MTEBI to the Moroccan context.

The use of the IRT to examine the psychometric properties of MTEBI is a key strength of this study. This approach has rarely been used in the past (54). Most researchers have preferred the CTT to test the psychometric properties of MTEBI. The combination of CTT and IRT adopted in this study is therefore innovative in the context of developing instruments to measure beliefs in the effectiveness of mathematics teaching. In line with expert recommendations (55, 56). This combination is particularly beneficial for producing high-quality instruments. By jointly applying CTT and IRT, we aim to propose best practices for the development of standardized instruments for measuring beliefs in the effectiveness of mathematics instruction, which may inspire the development of other psychological instruments.

Implications

These results validate the MTEBI's cross-cultural utility within the Moroccan teacher education framework and underscore the diagnostic advantages of IRT over traditional CTT metrics. This study demonstrates that researchers must supplement global reliability indices with conditional precision estimates (θ) to ensure instrument sensitivity across all latent trait levels. Practically, the MTEBI—specifically the PMTE subscale—functions as an effective diagnostic tool for identifying trainees with lower self-efficacy, facilitating targeted pedagogical interventions early in initial training. Conversely, extreme Outcome Expectancy (OE) scores should be interpreted with caution due to localized precision loss. Future refinements should prioritize rephrasing misaligned items (e.g., SE17) and adjusting

linguistic anchors to ensure semantic and structural equivalence across Arabic- and French-speaking contexts.

Limitations

Despite its methodological rigor, this study's generalizability is constrained by its regional scope. While the sample included trainees from Fez-Meknes and Oujda, it may not fully reflect the institutional diversity of all 12 Moroccan CRMEFs (57). Future research should employ multigroup confirmatory factor analysis (MGCFA) to test the factor invariance of the MTEBI across other regions, ensuring the stability of its structure within the broader national teacher training context (58).

Furthermore, the absence of external validation and Differential Item Functioning (DIF) analyses constrains the assessment of construct validity and measurement equity. Future research should incorporate established instruments, such as the Teachers' Sense of Efficacy Scale (TSES), and social desirability measures to strengthen convergent validity and mitigate self-report bias (48, 59–61). Additionally, verifying measurement invariance across gender, region, and educational background via robust DIF detection is essential (62, 63). Subsequent studies must investigate the underlying sources of potential item bias—including cultural nuances in verbal anchors or internship-related variance—to ensure the comparability of MTEBI scores across the diverse Moroccan educational landscape.

Furthermore, the absence of external validation and Differential Item Functioning (DIF) analyses constrains the assessment of construct validity and measurement equity. Future research should incorporate established instruments, such as the Teachers' Sense of Efficacy Scale (TSES), and social desirability measures to strengthen convergent validity and mitigate self-report bias (63). Additionally, verifying measurement invariance across gender, region, and educational background via robust DIF detection is essential (62). Subsequent studies must investigate the underlying sources of potential item bias—including cultural nuances in verbal anchors or internship-related variance, to ensure the comparability of MTEBI scores across the diverse Moroccan educational landscape.

While the Graded Response Model (GRM) was employed based on item ordinality, future research should incorporate comparative model fitting against other polytomous IRT models (e.g., Partial Credit or Rating Scale models) using information criteria (AIC/BIC) to ensure the most parsimonious fit (24). Furthermore, this study's cross-sectional design precludes the analysis of developmental dynamics in efficacy beliefs. Longitudinal research utilizing latent growth modeling is necessary to track individual trajectories and identify factors—such as gender, region, or discipline—that influence the evolution of MTEBI scores throughout the multi-year teacher training process (64, 65). Addressing these limitations will refine the MTEBI's psychometric robustness and enhance the evidence base for teacher education practices globally.

5. Conclusion

This study validates the Arabic MTEBI for Moroccan pre-service teachers using a synergistic CTT-IRT approach. CFA corroborated the original two-factor structure—personal efficacy and outcome expectancy—with robust fit indices and clear discriminant validity. While the Graded Response Model (GRM) proved optimal for outcome expectancy, the self-efficacy subscale exhibited localized precision limits requiring further structural refinement. Critically, IRT analysis transcended CTT by identifying conditional reliability across the latent continuum, providing diagnostic granularity for extreme scores that global indices overlook. These findings establish the MTEBI’s cross-cultural robustness and provide empirical benchmarks for assessing teacher efficacy beliefs in diverse linguistic and educational contexts.

Declarations

Acknowledgements

The authors would like to express their sincere gratitude to the preservice teachers for their time and valuable contributions in completing the research questionnaires. Special thanks are also extended to the trainers and the administration of the Regional Center for Education and Training Professions (CRMEF) for their essential logistical support and cooperation throughout the data collection process.

Authors’ contributions

AJ has been involved in conceiving and designing the study, acquiring data, analyzing and interpreting it, and drafting the manuscript. **JE** has contributed to the conception, performed statistical analysis, revised the manuscript, and designed the study, given final approval for the paper to be published, taken responsibility for the integrity of the work, and is designated as 'guarantor. **YH** has been involved in conceiving and designing the study, acquiring data, analyzing and interpreting data, and drafting the manuscript. **KS** has contributed to the conception, reviewed the manuscript, and designed the study. **AM** has conducted statistical analysis, conception, design, and data interpretation. **KE** has conducted statistical analysis, conception and design, interpreted the data. **AB** has contributed to conceptualization, data curation, methodology, investigation. **TA** has contributed to data acquisition and interpretation and drafted the manuscript.

Funding

There is no funding to report.

Data availability

The data may be made available on genuine demand with the corresponding author.

Ethics approval and consent to participate

This study obtained approval from the Research Ethics Committee of the Centre Régional des Métiers d'Éducation et de Formation, Fès Meknès (approval code: CRMEF-30-12-2024, 02 /24). All procedures performed in this study were in accordance with the ethical standards of the CRMEF's Research Ethics Committee and with the 1964 Helsinki declaration and its later amendments. This study is conducted in full compliance with the ethical principles governing educational research. Participation is strictly voluntary, with informed consent obtained from all individuals and the right to withdraw guaranteed. Confidentiality and anonymity are safeguarded, ensuring that no personal or institutional identities are disclosed. The research design prioritizes non-maleficence, avoiding any psychological, social, or academic harm, while promoting equity and fairness across all participant groups. Transparency of purpose and procedures is maintained throughout, and all data are handled with scientific integrity, free from fabrication or falsification. The study is guided by beneficence, aiming to contribute positively to educational practice and policy, and adheres to all relevant institutional, national, and international regulations.

Consent for publication

Written informed consent for publication was obtained.

Conflicts of Interest

The authors declare that they have no conflicts of interest. No financial, personal, political, or religious interests have influenced the research or the preparation of this manuscript.

Exclusivity

The manuscript is original and not published or under consideration elsewhere.

Authorship Verification

All listed co-authors meet established authorship criteria.

GenAI Disclosure

Generative AI was used only for text refinement. It did not influence the study's design, analysis, or interpretation. The authors reviewed and take full responsibility for the manuscript's content and integrity.

References

1. Alford K, Winters J. Comparing Self-Efficacy of Preservice and Inservice Teachers: What Factors Influence the Teacher Efficacy of Undergraduate and Graduate Students. *Critical Questions in Education*. 2025;16(2):128–44.

2. Bjerke AH, Xenofontos C. Teachers' self-efficacy in teaching mathematics: tracing possible changes from teacher education to professional practice. *Teachers and Teaching*. 2024 Jan 2;30(1):1–15. doi:10.1080/13540602.2023.2219982
3. Hoy AW, Spero RB. Changes in teacher efficacy during the early years of teaching: A comparison of four measures. *Teaching and teacher education*. 2005;21(4):343–56.
4. Bandura A. *Self-efficacy: The exercise of control*. Vol. 11. Freeman; 1997.
5. Chen S, Geesa RL, Izci B, Song H. Investigating Preservice Teachers' Science and Mathematics Teaching Efficacy, Challenges, and Support. *The Teacher Educator*. 2022 Jul 3;57(3):304–24. doi:10.1080/08878730.2021.2007560
6. Smith E, Smith RA, Price M, Hebert M, Powell S. Preservice Teachers' Experiences With, Beliefs About, Self-Efficacy In, and Instructional Practices for Mathematical Writing. *Investigations in Mathematics Learning*. 2025 Apr 3;17(2):138–55. doi:10.1080/19477503.2024.2398958
7. Berd-Gómez R, González-Martín AM, Santos FH. Unveiling the anxiety about teaching math in Spain: A psychometric study of MAST15 and MTAS19 instruments. *Acta Psychologica*. 2026 Mar 1;263:106306. doi:10.1016/j.actpsy.2026.106306
8. Cil O, Akcay AO. A mixed method quasi-experimental study on transforming preservice teachers' mathematics anxiety and teaching self-efficacy beliefs. *J Math Teacher Educ*. 2025 Feb 1;28(1):179–209. doi:10.1007/s10857-024-09632-8
9. Briley JS. The relationships among mathematics teaching efficacy, mathematics self-efficacy, and mathematical beliefs for elementary pre-service teachers. *Issues in the undergraduate mathematics preparation of school teachers* [Internet]. 2012 [cited 2026 Feb 5];5. Available from: <https://eric.ed.gov/?id=EJ990482>
10. Çiftçi SK, Karadağ E. Analysis of mathematics teachers' beliefs on mathematics instruction and teaching self-efficiencies within the scope of flow theory. *Annales Universitatis Paedagogicae Cracoviensis | Studia ad Didacticam Mathematicae Pertinentia*. 2019 Dec 5;11:73–87. doi:10.24917/20809751.11.5
11. Gullo G, Gentile A, Caci B, Alesi M. The role of teachers' conception of students' intelligence, self-efficacy and need frustration and satisfaction in shaping tendencies in teaching practices. *Teaching and Teacher Education*. 2025 Jul 1;160:105033. doi:10.1016/j.tate.2025.105033
12. Hourigan M, Leavy AM. Pre-service primary teachers' mathematics teaching efficacy on entry to initial teacher education. *European Journal of Mathematics and Science Education*. 2022;3(1):17–33.
13. Twohill A, NicMhuirí S, Harbison L, Karakolidis A. Primary Preservice Teachers' Mathematics Teaching Efficacy Beliefs: the Role Played by Mathematics Attainment, Educational Level, Preparedness to Teach, and Gender. *Int J of Sci and Math Educ*. 2023 Feb;21(2):601–22. doi:10.1007/s10763-022-10259-5
14. Segarra J, Julià C. Mathematics Teaching Efficacy Belief and Attitude of Pre-Service Teachers and Academic Achievement. *European Journal of Science and Mathematics Education*. 2022;10(1):1–

- 14.
15. Riggs IM, Enochs LG. Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education*. 1990;74(6):625–37. doi:10.1002/sce.3730740605
 16. Enochs LG, Smith PL, Huinker D. Establishing Factorial Validity of the Mathematics Teaching Efficacy Beliefs Instrument. *School Sci & Mathematics*. 2000 Apr;100(4):194–202. doi:10.1111/j.1949-8594.2000.tb17256.x
 17. Cakiroglu E. The teaching efficacy beliefs of pre-service teachers in the USA and Turkey. *Journal of Education for Teaching*. 2008 Feb;34(1):33–44. doi:10.1080/02607470701773457
 18. Alkhateeb HM. Internal Consistency Reliability and Validity of the Arabic Translation of the Mathematics Teaching Efficacy Beliefs Instrument. *Psychol Rep*. 2004 Jun;94(3):833–8. doi:10.2466/pr0.94.3.833-838
 19. Ryang. Development of the Mathematics Teaching Efficacy Beliefs Instrument Korean version for elementary preservice teachers. *Mathematics Education*. 2013;52(3):363–77.
 20. Koutsianou A, Emvalotis A. Greek pre-service primary teachers' efficacy beliefs in science and mathematics teaching: Initial adaptation of the STEBI-B and MTEBI instruments. *International Journal of Educational Methodology*. 2019;5(3):375–85.
 21. Mihajlović AM, Vulović NR, Milikić MP. The Teaching Efficacy of Preservice Mathematics Teachers: Research in the Republic of Serbia. *The New Educational Review*. 2020;61:63–75.
 22. Segarra J, Julià C, Valls C. Pre-Service Teachers' Belief About the Efficacy of Their Mathematics Teaching: A Case Study. *EUR J SCI MATH ED*. 2021 Sep 23;9(4):199–210. doi:10.30935/scimath/11236
 23. Hendricks W, Olawale BE, Saddiq K. An investigation of high school preservice teachers' self-efficacy in teaching mathematics. *Education Sciences*. 2024;14(11):1262.
 24. De Ayala RJ. *The Theory and Practice of Item Response Theory*. 2nd ed. 2022.
 25. Chen Y, Li X, Liu J, Ying Z. Item response theory—A statistical framework for educational and psychological measurement. *Statistical Science*. 2025;40(2):167–94.
 26. Van der Linden WJ. *Handbook of item response theory: Three volume set*. CRC Press; 2018.
 27. Wang C, Zhu R. Detecting uniform differential item functioning for continuous response computerized adaptive testing. *Applied Psychological Measurement*. 2024 Mar;48(1–2):18–37. doi:10.1177/01466216241227544
 28. Zein RA, Akhtar H. Getting started with the graded response model: An introduction and tutorial in R. *Int J Psychol*. 2025 Feb;60(1):e13265. doi:10.1002/ijop.13265
 29. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*. 1969;34(S1):1–97.
 30. Hair Jr. JF, Black WC, Babin BJ, Anderson RE. *Multivariate Data Analysis*. 8e édition. Andover, Hampshire: Cengage; 2019.
 31. Kline RB. *Principles and practice of structural equation modeling*. Guilford publications; 2023.

32. Omari M, Amaadour L, Zarrouq B, Al-Sheikh YA, El Asri A, Kriya S, et al. Evaluation of psychological distress is essential for patients with locally advanced breast cancer prior to neoadjuvant chemotherapy: baseline findings from cohort study. *BMC Women's Health*. 2023 Aug 23;23(1):445. doi:10.1186/s12905-023-02571-1
33. El-Ammari A, El Malki H, Moutawakkil SG, El Hilaly J, El Houari F, El Gnaoui S, et al. Validation of the Center for Epidemiologic Studies Depression Scale (CES-D) in a Moroccan sample with substance use disorder. *BMC Psychiatry*. 2023 Oct 6;23(1):723. doi:10.1186/s12888-023-05245-2
34. Watkins M. *A Step-by-Step Guide to Exploratory Factor Analysis with R and RStudio*. 2020. doi:https://doi.org/10.4324/9781003120001
35. Yong AG, Pearce S. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*. 2013;9(2):79–94.
36. Fornell C, Larcker DF. Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics. *Journal of Marketing Research*. 1981 Aug;18(3):382–8. doi:10.1177/002224378101800313
37. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 1999 Jan;6(1):1–55. doi:10.1080/10705519909540118
38. Ferrando PJ, Navarro-González D, Morales-Vives F. Linear and Nonlinear Indices of Score Accuracy and Item Effectiveness for Measures That Contain Locally Dependent Items. *Educational and Psychological Measurement*. 2025 Feb;85(1):60–81. doi:10.1177/00131644241257602
39. Quan Y, Wang C. Collapsing or not? A practical guide to handling sparse responses for polytomous items. *Methodology*. 2025;21(1):46–73.
40. Sanmuang S, Tuksino P, Chaipah K. Developing and Validating a Higher-Order Thinking and Innovation Competency Assessment: Psychometric Evidence for a Robust Adaptive Testing Item Bank. *TPM–Testing, Psychometrics, Methodology in Applied Psychology*. 2025;32(3):310–8.
41. Reeve B. Item Response Theory [IRT]. In: Maggino F, editor. *Encyclopedia of Quality of Life and Well-Being Research*. Cham: Springer International Publishing; 2024. p. 3689–97. doi:10.1007/978-3-031-17299-1_1556
42. Orlando M, Thissen D. Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*. 2003 Jul;27(4):289–98. doi:10.1177/0146621603027004004
43. Su S, Wang C, Weiss DJ. Performance of the S - χ^2 Statistic for the Multidimensional Graded Response Model. *Educational and Psychological Measurement*. 2021 Jun;81(3):491–522. doi:10.1177/0013164420958060
44. Carlson J. *Introduction to item response theory models and applications*. Routledge; 2020.
45. Matney G, Jackson JL, Panarach Y. Establishing Validity of the Thai Mathematics Teaching Efficacy Beliefs Instrument. *School Sci & Mathematics*. 2016 Mar;116(3):164–72. doi:10.1111/ssm.12165

46. Van Laar S, Braeken J. Understanding the Comparative Fit Index: It's All about the Base!. *Practical Assessment, Research & Evaluation*. 2021;26:26.
47. Bandura A. Self-efficacy: The exercise of control. Vol. 11. Freeman; 1997.
48. Tschannen-Moran M, Johnson D. Exploring literacy teachers' self-efficacy beliefs: Potential sources at play. *Teaching and Teacher Education*. 2011;27(4):751–61.
49. Safrudiannur. Measuring Teachers' Beliefs Quantitatively: Criticizing the Use of Likert Scale and Offering a New Approach [Internet]. Wiesbaden: Springer Fachmedien; 2020 [cited 2026 Apr 4]. (Kölner Beiträge zur Didaktik der Mathematik). Available from: <http://link.springer.com/10.1007/978-3-658-30023-4> doi:10.1007/978-3-658-30023-4
50. He J, Van de Vijver FJ, Fetvadjev VH, de Carmen Dominguez Espinosa A, Adams B, Alonso-Arbiol I, et al. On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures. *European Journal of Personality*. 2017;31(6):642–57.
51. Muraki E. A Generalized Partial Credit Model. In: Van Der Linden WJ, Hambleton RK, editors. *Handbook of Modern Item Response Theory* [Internet]. New York, NY: Springer New York; 1997 [cited 2026 Feb 5]. p. 153–64. Available from: http://link.springer.com/10.1007/978-1-4757-2691-6_9 doi:10.1007/978-1-4757-2691-6_9
52. Alkhateeb HM. Internal Consistency Reliability and Validity of the Arabic Translation of the Mathematics Teaching Efficacy Beliefs Instrument. *Psychol Rep*. 2004 Jun;94(3):833–8. doi:10.2466/pr0.94.3.833-838
53. Ryang D. Exploratory analysis of Korean elementary preservice teachers' mathematics teaching efficacy beliefs. *International Electronic Journal of Mathematics Education*. 2012;7(2):45–61.
54. Kieftenbeld V, Natesan P, Eddy C. An Item Response Theory Analysis of the Mathematics Teaching Efficacy Beliefs Instrument. *Journal of Psychoeducational Assessment*. 2011 Oct;29(5):443–54. doi:10.1177/0734282910391062
55. DeVellis RF, Thorpe CT. *Scale development: Theory and applications*. Sage publications; 2021.
56. Kyriazos TA, Stalikas A. Applied psychometrics: The steps of scale development and standardization process. *Psychology*. 2018;9(11):2531–60.
57. Marzoratti A, Evans TM. Why and how to collect representative study samples in educational neuroscience research. *Trends in neuroscience and education*. 2024;35:100231.
58. Sandoval-Ríos F, Cabezas-Orellana C, López-Núñez JA. Validation of a Spanish-language scale on data-driven decision-making in pre-service teachers. *Education Sciences*. 2025;15(7):789.
59. Lanz L, Thielmann I, Gerpott FH. Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*. 2022 Apr;90(2):203–21. doi:10.1111/jopy.12662
60. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*. 2016;41:71–90.

61. Crede M, Harms P. Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*. 2019;34(1):18–30.

62. Schneider L, Strobl C, Zeileis A, Debelak R. An R toolbox for score-based measurement invariance tests in IRT models. *Behav Res*. 2021 Dec 16;54(5):2101–13. doi:10.3758/s13428-021-01689-0

63. Li H, Hunter CV, Bialo JA. A Revisit of Zumbo’s Third Generation DIF: How Are We Doing in Language Testing? *Language Assessment Quarterly*. 2022 Jan 1;19(1):27–53. doi:10.1080/15434303.2021.1963253

64. Ma K, Dong J, McMaugh A, Cui Y. How pre-service teacher self-efficacy changes during the professional experience placement: a growth mixture model. *Aust Educ Res*. 2025 Apr;52(2):1367–87. doi:10.1007/s13384-024-00766-5

65. Yurt E. The relationships among pre-service teachers’ critical thinking disposition, self-efficacy, and creative thinking disposition in Turkey: a latent growth mediation model. *Curr Psychol*. 2025 Jan;44(1):85–102. doi:10.1007/s12144-024-07147-2

Figures

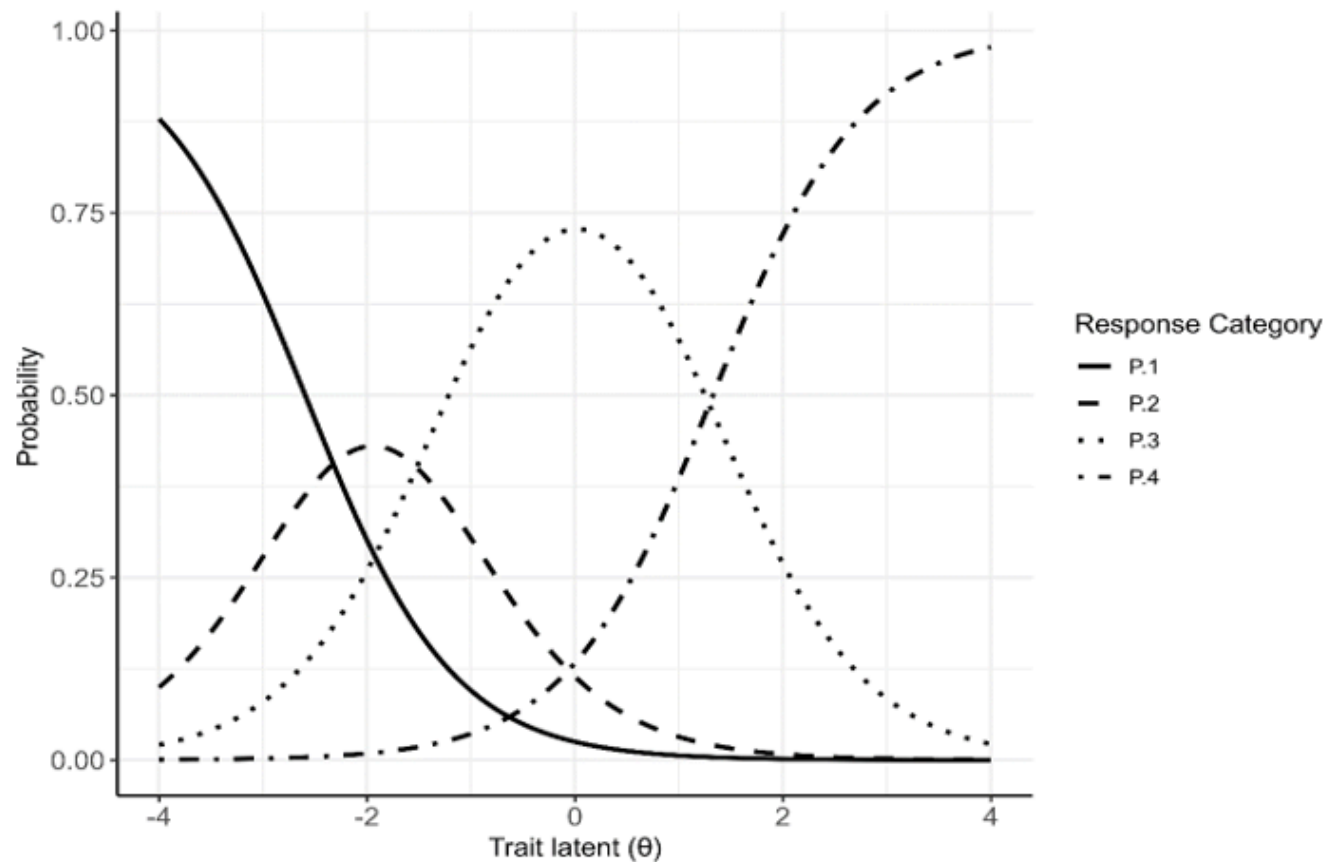


Figure 1

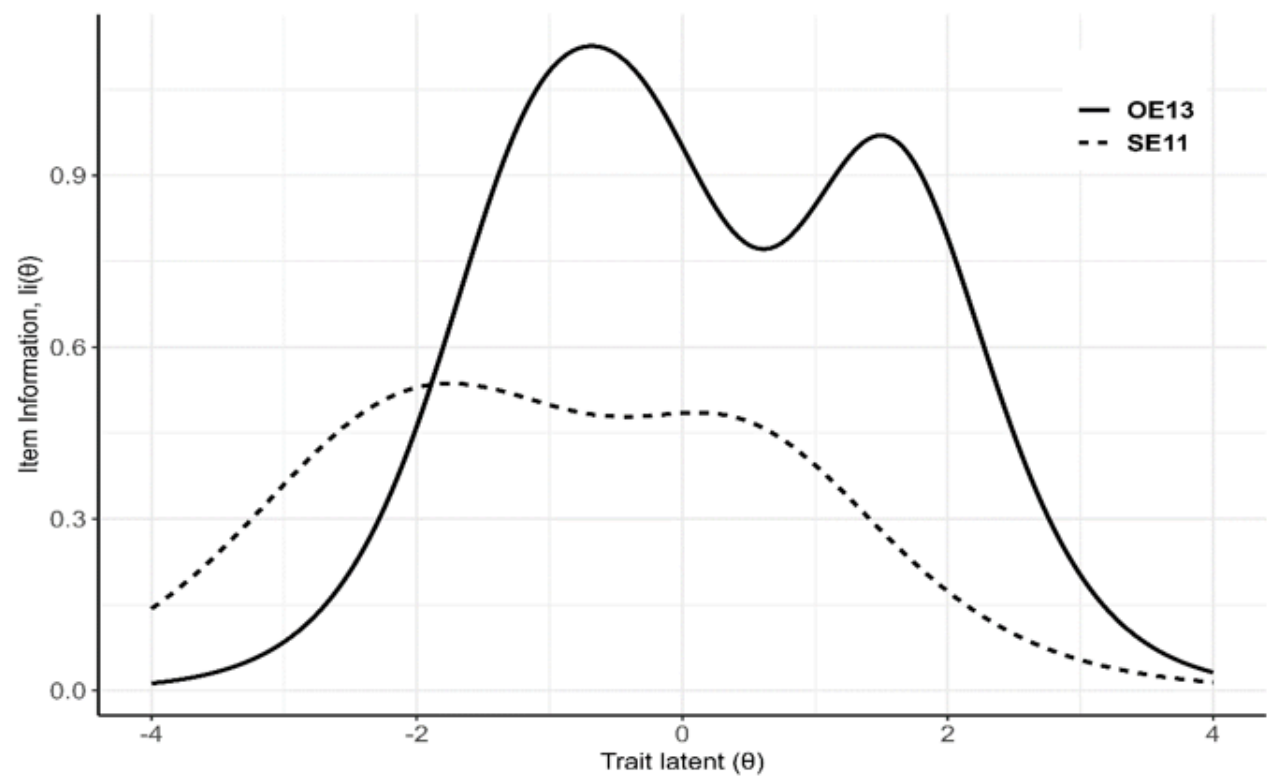


Figure 2

Item Information Curves for SE11 and OE13

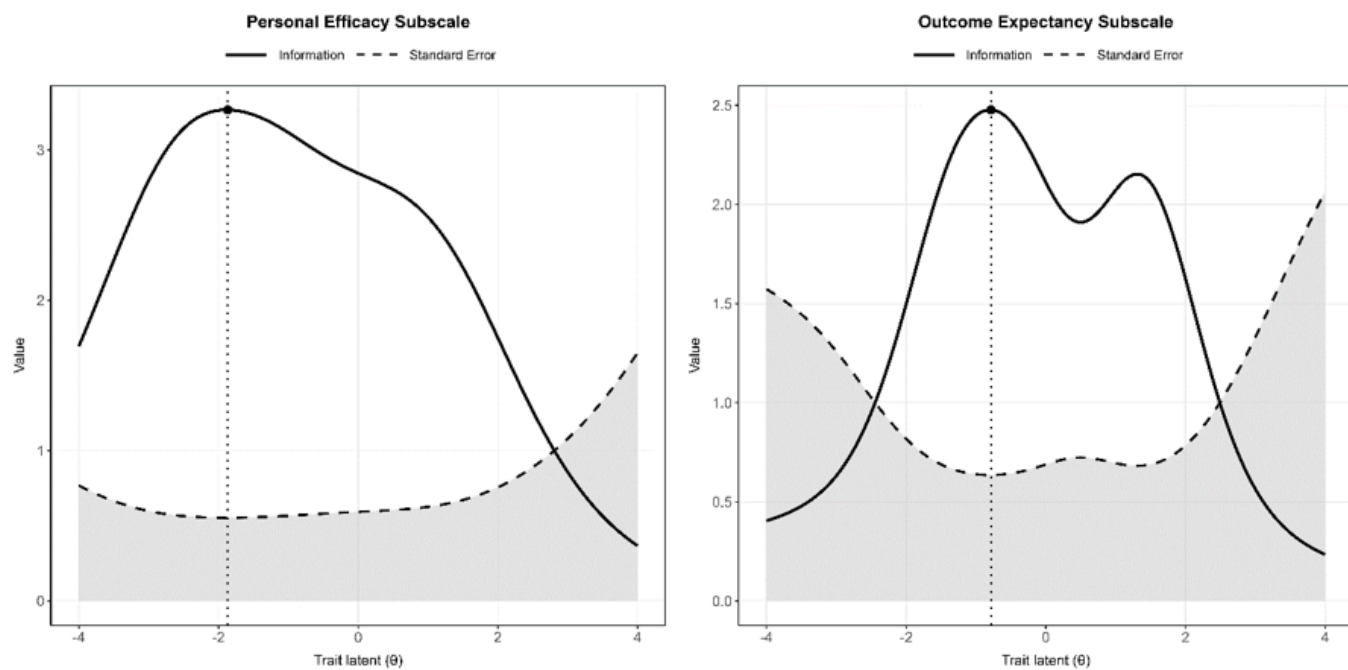


Figure 3

IRT Information Functions and Standard Errors (Comparison of Measurement Precision Across Subscales)

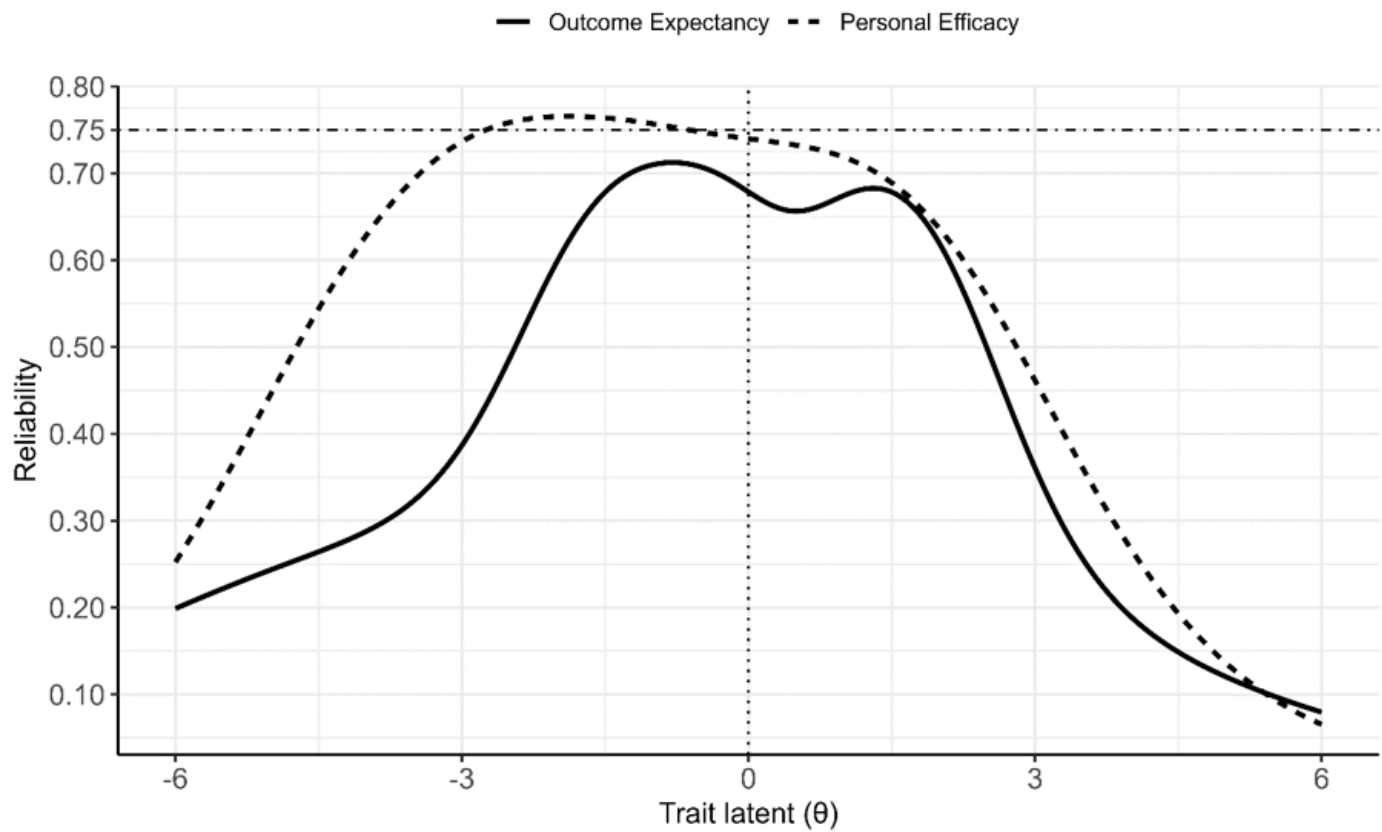


Figure 4

Reliability Curves Across Subscales