MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4101-4104

PSYCHOMETRIC VALIDATION OF THE POLICE SELECTION INSTRUMENT USING AN ORTHOGONAL MULTIDIMENSIONAL RASCH MODEL (MRM)

Yofran Hengki Ndoluanak[1], Muchlas Suseno[2], Ahmad Ridwan[3]
1,2,3 Doctoral Program in Educational Research and Evaluation, Universitas Negeri Jakarta (UNJ), Jakarta, Indonesia

ABSTRACT

Traditional assessment tools for specialized police units often collapse distinct competencies into a single composite score, obscuring the operationally critical differences among specialist roles. This study challenges that assumption by introducing a psychometric framework that treats counter-terrorism, bomb disposal, chemical-biological-radiological-nuclear (CBRN) threat management, and technical support as fundamentally independent measurement dimensions. A sixty-item selection instrument was developed for Detachment Gegana of the Indonesian National Police and validated using the Orthogonal Multidimensional Rasch Model (OMRM) with data from 321 active officer candidates across twelve regional units. A sequential validation framework integrated exploratory factor analysis, confirmatory factor analysis, model selection, item fit assessment, reliability estimation, and differential item functioning analysis. The orthogonal four-factor confirmatory model demonstrated excellent fit ($RMSEA = 0.005, CFI = 0.990, TLI = 0.990$), with a non-significant chi-square difference test ($\Delta\chi^2 = 0.834, df = 6, p = 0.991$) confirming that expertise in one domain carries virtually no predictive value for another. The Multidimensional Rasch Model was preferred over its two-parameter alternative based on substantially superior parsimony-adjusted fit ($\Delta BIC = 271.70$). All sixty items met acceptable fit criteria (RMSEA: $0.000–0.095$), item reliability was uniformly excellent across domains (range: $0.979–0.987$), and no item exhibited significant differential item functioning after false discovery rate correction. These findings establish that effective selection for specialized roles requires measuring each competency independently, offering a validated and practically actionable framework for evidence-based personnel assignment in elite law enforcement units.

Keywords: *multidimensional measurement, specialized competency assessment, personnel selection, item response theory, differential item functioning*

1. INTRODUCTION

Personnel selection for specialized law enforcement units represents one of the highest-stakes assessment contexts in applied measurement. Measurement error in this setting extends beyond individual misclassification: it may directly compromise operational effectiveness, public safety, and institutional integrity. These environments demand instruments that assess not a general ability but a clearly defined set of operationally grounded competency domains, each requiring distinct technical knowledge, tactical judgment, and situational reasoning under conditions of risk (McDaniel et al., 2001; Sackett et al., 2022).

Despite growing recognition of classical test theory's limitations, many specialized recruitment programs continue to use total-score approaches that assume a single latent trait. When an instrument is designed to measure conceptually distinct, operationally non-interchangeable domains, a composite score conflates performance across those domains and makes it impossible to identify domain-specific strengths and weaknesses. Furthermore, item difficulty and person ability estimates under classical frameworks are sample-dependent, limiting the generalizability of psychometric conclusions across cohorts (T. G. Bond & Fox, 2015; Hambleton & Jones, 1993).

Rasch measurement theory offers a compelling alternative by producing sample-invariant estimates of item difficulty and person ability on a common logit scale. When an instrument targets multiple independent domains, Multidimensional Rasch Models—specifically the Orthogonal variant—allow each item to load on a single designated dimension while providing independent latent ability estimates per domain. The orthogonality assumption posits that candidate performance in one operational specialty carries no predictive information about performance in another, which is theoretically and empirically defensible when specializations are acquired through distinct training pipelines (Adams et al., 1997; Reckase, 2009).

The present study applies the OMRM to validate a sixty-item selection instrument developed for the 2025 recruitment cycle of Detachment Gegana, the elite tactical unit of the Indonesian National Mobile Police Brigade. Three primary objectives guided this study: (1) to examine whether the instrument's dimensional structure is consistent with a theoretically specified four-factor orthogonal model; (2) to evaluate item-level and global model fit; and (3) to assess reliability, separation, and person-item targeting. The central research question is: Does the sixty-item instrument demonstrate adequate multidimensional construct validity and measurement quality under the OMRM?

2. LITERATURE REVIEW

2.1 Rasch Measurement Theory

Rasch measurement theory (Rasch, 1977; B. Wright & Stone, 1979) departs from classical test theory by estimating ability on a continuous logit scale independent of the particular items administered. This sample invariance property is fundamental for selection contexts requiring equitable assessment across cohorts. Item difficulty and person ability are estimated jointly via maximum likelihood, and the fit of observed data to the model serves as an empirical test of construct validity (T. G. Bond & Fox, 2015; M. Wilson, 2004). Within this framework, unidimensionality requires that item response residuals, after accounting for the latent ability, are mutually independent; violations indicate secondary constructs that distort person estimates.

2.2 Multidimensional Rasch Models

Adams et al. (1997) formalized multidimensional Rasch extensions through the Multidimensional Random Coefficients Multinomial Logit (MRCML) model, which estimates ability on multiple dimensions while maintaining Rasch-consistent constraints on item discrimination. When items are assigned to a single dimension in a simple-structure loading matrix, the model yields independent ability estimates per domain while preserving Rasch parsimony and interpretability (Wu, 2007). Research consistently demonstrates that multidimensional models produce more accurate person estimates, reduce misfit, and improve score interpretability relative to forced unidimensional solutions (de la Torre & Patz, 2005; Yao & Boughton, 2007).

2.3 Orthogonal Multidimensional Structure

The orthogonal model specifies that measured dimensions are statistically uncorrelated, appropriate when each operational specialization requires a distinct body of knowledge acquired through independent training. Validation of the orthogonality assumption is typically conducted through confirmatory factor analysis, comparing correlated and uncorrelated factor models using goodness-of-fit indices and the Satorra-Bentler scaled chi-square difference test (Hu & Bentler, 1999; Reckase, 2009)

3. METHODOLOGY

3.1 Participants

Participants were 321 active-duty police officers (male: $n = 289$, 90%; female: $n = 32$, 10%) who applied for Detachment Gegana selection during the 2025 recruitment cycle. All held a minimum of three years of prior service. Mean age was 25.1 years ($SD = 1.55$) and mean service length was 4.06 years ($SD = 1.28$). Educational attainment comprised secondary school (48.6%), undergraduate degree (44.2%), and diploma qualifications (7.2%). Candidates were drawn from twelve regional mobile police brigade units, with the largest contributions from the

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4101-4104

Metropolitan Jakarta (19.9%), East Java (13.4%), and West Java (12.8%) units, supporting the generalizability of psychometric findings across the target population.

3.2 Instrument

The selection instrument consists of sixty dichotomously scored multiple-choice items distributed equally across four operational competency domains of fifteen items each: Counter-Terror Operations (items 1–15), Bomb Disposal Operations (items 16–30), Specialized Technical Support Operations (items 31–45), and CBRN Threat Management (items 46–60). Items employed three formats: situational judgment scenarios, technical knowledge items, and risk-based screening items. Content validity was established through expert panel review involving operational practitioners, measurement specialists, a psychometrician, and a national security education expert. Content validity indices were calculated following Lawshe (1975) and revised by F. R. Wilson et al. (2012), yielding an overall instrument CVI of 0.883, exceeding the recommended threshold of 0.80 (Polit & Beck, 2006).

3.3 Analytical Model

The OMRM models the probability that person j responds correctly to item i assigned to dimension d as:

$$P(X_{ij} = 1|\theta_{jd}, \beta_i) = \frac{\exp(\theta_{jd} - \beta_i)}{1 + \exp(\theta_{jd} - \beta_i)}$$

where θjd denotes person j's latent ability on dimension d and β_i is item i's difficulty on the logit scale. The orthogonality constraint, $Cov(\theta_{jd}, \theta_{jd'}) = 0$ for all $d \neq d'$, specifies that ability on any dimension carries no predictive information about ability on any other (Reckase, 2009; Wu, 2007). Parameters were estimated using marginal maximum likelihood with the quasi-Monte Carlo expectation-maximization algorithm (QMCEM) in the mirt package (Chalmers, 2012).

3.4 Data Analysis

All analyses were performed in R 4.4.2. The five-stage sequential framework proceeded as follows: (1) Exploratory Factor Analysis using polychoric correlations and minimum-residual extraction (psych package); (2) Confirmatory Factor Analysis comparing correlated and orthogonal four-factor models with WLSMV estimation (lavaan package); (3) Model selection comparing the Multidimensional Rasch Model against the Multidimensional 2PL using information criteria, likelihood ratio testing, and M2 absolute fit statistics (mirt package); (4) Item-level fit assessment using the S-X2 statistic with RMSEA per item (mirt package); and (5) Reliability, person separation, person-item targeting, and differential item functioning analysis using Wald tests with Benjamini-Hochberg false discovery rate correction (mirt package). Each stage was conditioned on evidence from the preceding stage, ensuring that conclusions regarding model fit, item quality, and reliability were built on empirically verified foundations.

4. RESULTS

4.1 Dimensional Structure

Exploratory factor analysis on the sixty-item polychoric correlation matrix yielded four dominant eigenvalues ($4.331, 3.971, 3.506, 3.383$) that substantially exceeded the fifth (1.190), producing a clear scree inflection after Factor 4. Although the Kaiser criterion suggested six factors and parallel analysis indicated sixteen, both outcomes were attributable to the known behavior of these procedures with binary data and polychoric correlations rather than to genuine additional dimensionality (Cho et al., 2009; Green et al., 2016). The four-factor solution produced interpretable simple structure: all items loaded substantially on their theoretically designated factors (loading range: $0.310–0.692$) with cross-loadings below 0.30. The four factors collectively explained 26.7% of item response variance, with approximately equal contributions across domains ($MR1$: 7.2%; $MR2$: 6.8%; $MR3$: 6.4%; $MR4$: 6.2%), confirming balanced dimensionality consistent with the equal-item-per-domain instrument design.

Confirmatory factor analysis compared a correlated four-factor model and an orthogonal four-factor model estimated with the WLSMV estimator. Table 1 presents complete fit indices. The orthogonal model produced excellent fit (scaled RMSEA = 0.005, 90% CI [0.000, 0.014]; scaled CFI = 0.990; TLI = 0.990; SRMR = 0.096) and was marginally superior to the correlated model on all scaled indices. All six interfactor correlations in the correlated model were negligible and non-significant (maximum $|r| = 0.049$; all $p > 0.51$). The Satorra-Bentler scaled chi-square difference test confirmed that constraining all interfactor covariances to zero produced no significant deterioration in fit ($\Delta\chi^2 = 0.834, df = 6, p = 0.991$), providing full empirical justification for the orthogonal specification.

Table 1. CFA Four-Factor Model Fit Comparison

Fit Index	Correlated Model	Orthogonal Model	Criterion
Chi-square (Standard)	1649.49 (df=1704, p=0.824)	1657.04 (df=1710, p=0.817)	p > 0.05
Chi-square (Scaled)	1736.12 (p=0.288)	1721.59 (p=0.417)	p > 0.05
RMSEA (Scaled)	0.008 [0.000, 0.016]	0.005 [0.000, 0.014]	< 0.05
CFI (Scaled)	0.973	0.990	> 0.95
TLI (Scaled)	0.972	0.990	> 0.95
SRMR	0.096	0.096	< 0.10
Δχ² (df=6)	—	0.834 (p=0.991)	p > 0.05

Note. Estimation: WLSMV. Scaled statistics employ Satorra-Bentler correction.

4.2 Model Selection

A formal comparison between the Multidimensional Rasch Model (MRM) and the Multidimensional Two-Parameter Logistic model (M2PL) was conducted using both relative and absolute fit criteria. Table 2 summarizes the key model comparison results. The MRM was favored over the M2PL across all four information criteria, with the BIC difference of 271.70 substantially exceeding the threshold of 10 units for very strong evidence in favor of the more parsimonious model (Raftery, 1995). The likelihood ratio test confirmed that the M2PL's 56 additional discrimination parameters yielded no significant improvement in fit ($\chi^2 = 51.498, df = 56, p = 0.646$). Absolute fit via the M2 statistic further supported MRM superiority (RMSEA = 0.008 relative to 0.010; CFI = 0.986 relative to 0.980; TLI = 0.986 relative to 0.979). The MRM was retained as the primary analytical framework.

Table 2. Model Fit Comparison

Index	MRM	M2PL	Difference	Preferred
AIC	21908.44	21968.94	60.50	MRM
BIC	22172.44	22444.14	271.70	MRM
RMSEA (M2)	0.008	0.010	—	MRM
CFI (M2)	0.986	0.980	—	MRM
TLI (M2)	0.986	0.979	—	MRM
LRT p-value	—	0.646	—	MRM

Note. LRT = Likelihood Ratio Test (df = 56). BIC difference of 271.70 exceeds the threshold of 10 for very strong evidence (Raftery, 1995).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4101-4104

4.3 Item Fit Statistics

Item-level fit was assessed using the S-X2 statistic (Orlando & Thissen, 2000, 2003), with an item classified as fitting when its associated RMSEA fell below 0.10. All sixty items met this criterion across all four domains (Table 3). The three items with RMSEA values nearest the threshold—Item 36 (0.095), Item 18 (0.081), and Item 25 (0.079)—reflect the known behavior of S-X2 at extreme difficulty levels, where restricted response distributions amplify sensitivity to minor irregularities without substantively threatening construct validity.

Table 3. Item-Level Fit Statistics (S-X2, OMRM)

Domain	Items	RMSEA Range	Fit Decision
Counter-Terror Operations	1–15	0.029–0.076	15 Fit / 0 Misfit
Bomb Disposal Operations	16–30	0.000–0.081	15 Fit / 0 Misfit
Technical Support Operations	31–45	0.000–0.095	15 Fit / 0 Misfit
CBRN Threat Management	46–60	0.000–0.074	15 Fit / 0 Misfit
Full Instrument	**1–60**	**0.000–0.095**	**60 Fit / 0 Misfit**

Note. Fit criterion: RMSEA < 0.10 (Kang & Chen, 2008; Orlando & Thissen, 2000). CBRN = Chemical-Biological-Radiological-Nuclear.

4.4 Reliability and Measurement Precision

Person and item reliability indices are presented in Table 4. Person reliability ranged from 0.699 to 0.732 across domains, with all values approaching or meeting the 0.70 minimum threshold for applied measurement (T. Bond et al., 2020). These modest values are consistent with the known constraint that person reliability in multidimensional Rasch frameworks is bounded by subscale length; fifteen items per dimension represents a deliberate design trade-off between multidimensional breadth and subscale depth. Item reliability was uniformly excellent across all domains (range: $0.979–0.987$), confirming that the sample of 321 candidates produced stable, replicable item difficulty calibrations far exceeding the 0.90 criterion for high-quality item calibration (Engelhard Jr., 2013; B. Drake. Wright & Masters, 1982). Test information functions peaked near $\theta = 0$ across all four domains, with standard error of measurement rising steeply beyond $\theta = \pm 2$, confirming that measurement precision was concentrated precisely in the ability range most relevant for pass-fail selection decisions.

Table 4. Person and Item Reliability Under the OMRM

Domain	Items	Person Reliability	Item Reliability
Counter-Terror Operations	1–15	0.717	0.981
Bomb Disposal Operations	16–30	0.699	0.987
Technical Support Operations	31–45	0.732	0.979
CBRN Threat Management	46–60	0.713	0.983
Full Instrument	**1–60**	**0.715**	**0.984**

Note. Person reliability computed via empirical_rxx() on EAP factor scores. Item reliability = true-score variance ratio (B. Drake. Wright & Masters, 1982).

4.5 Differential Item Functioning

Because available subgroup variables (gender, education, regional unit) yielded cell sizes insufficient for stable multigroup IRT estimation, DIF was assessed using the odd-even split method, which assigns candidates to two groups by sequential row index (Group 1: n = 161 odd; Group 2: n = 160 even), providing a principled internal consistency check for item stability (Linacre, 2006). Wald tests were conducted within a multiple-group OMRM framework, with false discovery rate correction applied to all sixty simultaneous tests (Benjamini & Hochberg, 1995). Two items produced nominally significant unadjusted p-values (Item 43, $p = 0.043$; Item 52, $p = 0.049$); however, both yielded FDR-adjusted p-values of 0.662, consistent with chance occurrence under the null hypothesis. All sixty items were classified as DIF-free after adjustment (adjusted p-values: $0.662–0.990$), confirming measurement invariance across independent candidate subsamples.

4.6 Person-Item Targeting and Candidate Classification

Wright Map analysis confirmed that the bulk of candidate abilities and item difficulties overlapped between −1 and +1 logits across all four domains, indicating adequate instrument targeting for the central ability range where the majority of selection decisions are concentrated. The full-instrument difficulty range spanned −2.128 to 3.088 logits, with 43.3% of items classified as Moderate, 26.7% as Easy, 18.3% as Difficult, 6.7% as Very Difficult, and 5.0% as Very Easy, producing an appropriate difficulty distribution for a minimum-competency-threshold selection context. Candidate suitability classifications derived from domain-specific θ estimates showed the Versatile category as modal across all domains (33.3–38.9%), with 28.6–35.8% of candidates classified as Recommended or Highly Recommended and 28.8–32.4% classified as Not Recommended or Strongly Not Recommended, providing operationally actionable profiling information for specialist unit assignment.

5. DISCUSSION

This study successfully developed and validated a multidimensional operational competency selection instrument for Detachment Gegana using the OMRM. The confirmatory orthogonal four-factor model demonstrated excellent fit and all interfactor correlations were empirically negligible, establishing that counter-terror, bomb disposal, technical support, and CBRN competencies constitute genuinely independent knowledge structures rather than surface expressions of a single general operational ability. This finding has direct implications for construct validity: it means that a candidate who excels in CBRN decontamination procedures may possess no advantage—and no disadvantage—in explosive ordnance render-safe operations, validating the theoretical premise that these specializations require domain-specific training and assessment.The preference for the MRM over the M2PL is consistent with the broader psychometric literature demonstrating that equal-discrimination Rasch models provide adequate fit for specialized knowledge-based assessments where items are drawn from well-defined, homogeneous content domains (T. Bond et al., 2020; Reckase, 2009). The sample invariance and parsimony properties of the Rasch model are particularly valuable in a high-stakes selection context where score comparability across candidate cohorts administered at different time points is a measurement priority. Universal item fit, excellent item reliability, and absence of DIF collectively confirm that the instrument is stable, fair, and replicable.From a practical standpoint, the validated OMRM framework enables domain-specific latent ability scores to be used directly as selection criteria, allowing decision-makers to evaluate candidates on their specific readiness profiles across the four specialist domains rather than on a single composite. The five-category suitability classification provides an operationally intuitive profiling tool that translates psychometric scores into actionable selection recommendations without requiring end-users to possess specialist measurement expertise.

Three limitations qualify these conclusions. First, person reliability values at the domain level (0.699–0.732) fall below the 0.80 threshold commonly recommended for high-stakes individual decisions. Extending each subscale from fifteen to twenty or twenty-five items in future iterations would substantially improve precision at the extremes of the ability distribution, as demonstrated by the test information functions. Second, the sample of 321 candidates represents a single recruitment cohort, and the stability of item calibrations across future cohorts should be monitored through periodic equating studies. Third, DIF analysis against substantively meaningful subgroup variables—educational background, years of service, regional unit—should be conducted when subgroup sample sizes in subsequent administrations are sufficient for stable multigroup estimation.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 4101-4104

6. CONCLUSIONS

This study provides the first IRT-based psychometric validation of a specialized competency selection instrument for an elite Indonesian counter-terrorism unit. The sixty-item OMRM-validated instrument demonstrated that the four operational competency domains function as genuinely orthogonal latent constructs, that the Multidimensional Rasch Model is empirically preferred over less constrained alternatives, that all items fit the model, that item reliability is uniformly excellent, and that the instrument is free of differential item functioning. Person-item targeting analysis confirmed adequate coverage of the ability range most relevant for selection decisions. Together, these findings establish both a validated measurement tool and a methodological template for evidence-based personnel selection in comparable high-stakes law enforcement contexts. Future research should prioritize subscale expansion to improve person reliability, DIF analysis against meaningful demographic subgroups in larger samples, and periodic equating studies to ensure the long-term psychometric integrity of the instrument across successive Detachment Gegana recruitment cycles.

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model : Fundamental Measurement in the Human Sciences* (3rd ed., Vol. 11). Routledge: Taylor & Francis. https://doi.org/10.4324/9781315814698

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th Ed.). Routledge. https://doi.org/10.4324/9780429030499

Chalmers, R. P. (2012). mirt : A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the Parallel Analysis Procedure With Polychoric Correlations. *Educational and Psychological Measurement*, *69*(5), 748–759. https://doi.org/10.1177/0013164409332229

de la Torre, J., & Patz, R. J. (2005). Making the Most of What We Have: A Practical Application of Multidimensional Item Response Theory in Test Scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295–311. https://doi.org/10.3102/10769986030003295

Engelhard Jr., G. (2013). *Invariant Measurement*. Routledge. https://doi.org/10.4324/9780203073636

Green, S. B., Redell, N., Thompson, M. S., & Levy, R. (2016). Accuracy of Revised and Traditional Parallel Analyses for Assessing Dimensionality with Binary Data. *Educational and Psychological Measurement*, *76*(1), 5–21. https://doi.org/10.1177/0013164415581898

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kang, T., & Chen, T. T. (2008). Performance of the Generalized S-X^2 Item Fit Index for Polytomous IRT Models. *Journal of Educational Measurement*, *45*(4), 391–406. http://www.jstor.org/stable/20461906

Lawshe, C. H. (1975). A quantitative approach to content validity [1]. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. (2006). A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Programs. In *www.winsteps.com*.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 730–740. https://doi.org/10.1037/0021-9010.86.4.730

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Polit, & Beck. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res Nurs Health*, *29*(5), 489–497.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063

Rasch, G. (1977). On Specific Objectivity an Attempt at Formalizing the Request for Generality and validity of Scientific Statements. *Danish Yearbook of Philosophy*, *14*(1), 58–94. https://doi.org/10.1163/24689300-01401006

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer New York. https://doi.org/10.1007/978-0-387-89976-3

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040–2068. https://doi.org/10.1037/apl0000994

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the Critical Values for Lawshe's Content Validity Ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210. https://doi.org/10.1177/0748175612440286

Wilson, M. (2004). *Constructing Measures: An Item Response Modeling Approach* (1st Ed.). Routledge. https://doi.org/10.4324/9781410611697

Wright, B. Drake., & Masters, G. N. . (1982). *Rating scale analysis*. Mesa Press.

Wright, B., & Stone, M. (1979). *Best Test Design Rasch Measurement*. MESA Press.

Wu, M. L. . (2007). *ACER ConQuest version 2.0 : generalised item response modelling software*. ACER Press.

Yao, L., & Boughton, K. A. (2007). A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification. *Applied Psychological Measurement*, *31*(2), 83–105. https://doi.org/10.1177/0146621606291559