Volume 5, No. 11 November 2026 - (1350-1361)
p-ISSN 2980-4868 | e-ISSN 2980-4841
https://ajesh.ph/index.php/gp

Calibration and Equalization of Computer Adaptive Test Scores for Assessment Centers Using The Deming Regression Method in The Public Works Adaptive Test (ATPU)

Rud Nokeo[1]*, Fieghie Fadila[2], Gayatri Indah Puspitorini[3]
Universitas Gadjah Mada, Indonesia[1]
Universitas Katolik Indonesia Atma Jaya, Indonesia[2]
Universitas Airlangga, Indonesia[3]
Email: nokeorud@gmail.com[1]*, fieghiefadila@gmail.com[2], puspitorinigayatri@gmail.com[3]

Keywords:
Assessment Center; Computer Adaptive Test; Calibration Adaptive Test of Public Works (ATPU); Equalization of Scores; Deming Regression

ABSTRACT

The development of computer adaptive testing (CAT) through the Adaptive Public Works Test (ATPU) offers improved assessment efficiency through adaptive testing algorithms. However, differences in assessment stimulus formats may introduce systematic measurement bias, potentially resulting in more stringent competency ratings than those obtained through direct behavioral observation. This study aimed to identify systematic bias in raw CAT scores, develop linear score-equating functions using Deming regression across nine competency dimensions, and evaluate the validity and equivalence of the calibrated scores using an independent validation sample. A quantitative correlational and score-equating research design was employed, using competency assessment data divided into model-development and independent validation samples. Data analysis included Pearson and Spearman correlation analyses, estimation of Deming regression parameters under the assumption of equal measurement error variances, evaluation of continuous prediction accuracy, two one-sided tests (TOST) for statistical equivalence, and assessment of categorical classification agreement. The findings indicated that although raw CAT scores exhibited a very strong correlation with the reference assessment scores, they demonstrated systematic negative bias, resulting in the underestimation of participants' competency levels. The application of Deming regression-based calibration substantially reduced systematic bias, improved agreement as measured by the concordance correlation coefficient, and minimized prediction errors. Furthermore, the calibrated scores demonstrated statistical equivalence within the predefined equivalence margins and improved consistency in competency-level classification. These findings suggest that Deming regression provides a promising approach to improving score comparability and classification consistency between CAT-based assessments and direct behavioral assessments in assessment center settings.

INTRODUCTION

Human resource (HR) competency assessment plays a strategic role in ensuring objectivity in employee placement, promotion, talent management, and career development within merit-based organizational systems (Boyatzis, 2008; Spencer & Spencer, 1993). For decades, the assessment center (AC) method has been widely recognized as a rigorous approach to evaluating

managerial and sociocultural competencies because of its ability to reduce subjective bias through multiple behavioral simulations and observations conducted by trained assessors (International Taskforce on Assessment Center Guidelines, 2015; Thornton & Rupp, 2006). The integration of behavioral evidence from assessment exercises, including in-basket exercises, leaderless group discussions (LGDs), role-playing simulations, and behavioral event interviews, has demonstrated predictive validity for subsequent workplace performance (Arthur et al., 2003; Hermelin et al., 2007). Through assessor integration discussions, behavioral indicators are systematically evaluated to produce comprehensive and defensible estimates of individual competency levels (Williams et al., 2024). Despite its established psychometric advantages, conventional assessment center implementation faces substantial operational constraints, including high logistical costs, limited availability of certified assessors, and lengthy assessment procedures. These structural limitations hinder the rapid and efficient application of conventional assessment center methods, particularly when organizations must conduct large-scale competency assessments across extensive employee populations.

With the rapid digital transformation of modern psychometrics, computer adaptive testing (CAT), implemented through the Adaptive Public Works Test (ATPU), has emerged as a promising alternative for addressing these operational challenges (van der Linden & Glas, 2010; Magis et al., 2017). Grounded in item response theory (IRT), CAT employs adaptive algorithms that dynamically adjust item difficulty based on examinees' previous responses and updated estimates of their ability levels (Embretson & Reise, 2000; Lord, 1980). Adaptive item selection has demonstrated the potential to substantially reduce testing duration compared with conventional fixed-form assessments while maintaining measurement precision and supporting test security through controlled item exposure (Wainer et al., 2000; Weiss, 2011). However, measurement and test equating theory emphasizes that numerical scores generated by instruments employing different stimulus formats and administration procedures cannot automatically be considered interchangeable on their original measurement scales (Angoff, 1984; Kolen & Brennan, 2014). Differences between algorithm-driven item responses in CAT and complex behavioral observations in assessment centers may introduce systematic differences in measurement difficulty, scoring characteristics, and competency estimates. Consequently, directly substituting raw CAT scores for established assessment center scores without empirical calibration may compromise score comparability and introduce systematic bias into personnel decisions (Dorans et al., 2010; González & Wiberg, 2017).

The need for score calibration was further supported by preliminary diagnostic findings involving nine managerial and sociocultural competency dimensions assessed among Indonesian civil servants, namely Integrity, Cooperation, Communication, Results Orientation, Public Service, Development of Self and Others, Managing Change, Decision-Making, and National Cohesion. Preliminary analyses indicated strong Pearson product-moment correlations and Spearman rank-order correlations between raw CAT scores and assessment center scores. Nevertheless, systematic negative differences were observed across the nine competency dimensions, with CAT scores generally producing lower competency estimates than their

assessment center counterparts. These discrepancies resulted in some participants being assigned to lower competency classifications, potentially affecting the consistency and fairness of personnel decisions. Conventional score-linking approaches frequently employ ordinary least squares (OLS) regression, which assumes that predictor variables are measured without error. However, both CAT and assessment center scores are subject to measurement error, potentially violating this assumption and producing biased regression estimates (Deming, 1943; Linnet, 1993). In particular, measurement error in the predictor variable may introduce regression dilution, resulting in attenuated slope estimates and reduced calibration accuracy. This methodological limitation highlights the need for calibration approaches that explicitly account for measurement errors in both instruments. Accordingly, this study employed Deming regression, an errors-in-variables regression method that incorporates measurement error variances in both score variables, to establish linear score-linking functions across nine competency dimensions. The study aimed to examine the initial correlations and systematic differences between raw CAT and assessment center scores, estimate competency-specific Deming regression calibration equations, evaluate the accuracy and statistical equivalence of calibrated scores using independent validation samples, and assess improvements in competency-level classification agreement. By integrating measurement error modeling with independent validation, the study sought to provide empirical evidence supporting more accurate, comparable, and consistent competency assessments within merit-based human resource management systems.

METHOD

This study employed a quantitative approach with a correlational design and psychometric score calibration and linking procedures (Kolen & Brennan, 2014). The research was conducted within the competency assessment system for Indonesia's State Civil Apparatus (ASN), which assessed managerial and sociocultural competencies using conventional assessment center (AC) methods and computer adaptive testing (CAT). The research procedures comprised data integrity verification, preliminary diagnostic analysis, stratified random partitioning, Deming regression model development, and independent validation. The study population consisted of assessment participants who completed both assessment methods during the same evaluation period. Total sampling was applied to eligible records in the assessment database, yielding 391 participants (N = 391) with complete paired scores, no missing values, and no duplicate participant records.

The study utilized paired competency scores obtained from AC and CAT assessments across nine competency dimensions: Integrity, Cooperation, Communication, Results Orientation, Public Service, Development of Self and Others, Managing Change, Decision-Making, and National Cohesion. The AC method integrated multiple behavioral simulations and assessor evaluations to generate continuous competency scores and categorical competency levels on a five-point scale. The CAT instrument employed adaptive testing algorithms to estimate participants' competency levels across the same nine dimensions, producing continuous numerical scores and corresponding categorical classifications. The dataset contained paired scores from both instruments, including continuous measurements and categorical competency levels. To evaluate model generalizability

and reduce the risk of overfitting, the dataset was randomly divided using a stratified sampling procedure into a training set (n = 261) for parameter estimation and an independent validation set (n = 130) for assessing calibration performance.

Data processing and statistical analyses were performed using Python and relevant statistical libraries to ensure analytical reproducibility. Preliminary analyses included descriptive statistics, Pearson correlation coefficients, Spearman rank correlation coefficients, absolute-agreement intraclass correlation coefficients (ICC), Lin's concordance correlation coefficient (CCC), and mean paired differences to identify initial associations, agreement, and systematic bias between CAT and AC scores (Lin, 1989). Subsequently, competency-specific linear calibration equations were estimated from the training dataset using Deming regression with an assumed measurement error variance ratio of $\lambda = 1.0$, yielding intercept and slope parameters for each competency dimension (Deming, 1943; Passing & Bablok, 1983). The resulting equations were applied to the independent validation dataset to generate calibrated CAT scores on the AC reference scale.

Calibration performance was evaluated by comparing predicted and observed AC scores using mean absolute error (MAE), root mean squared error (RMSE), and Bland–Altman limits of agreement (Bland & Altman, 1986; Chai & Draxler, 2014). Statistical equivalence was assessed using the two one-sided tests (TOST) procedure at a significance level of $\alpha = 0.05$, with equivalence determined according to predefined equivalence margins (Lakens, 2017; Schuirmann, 1987). Agreement in categorical competency classifications was evaluated using quadratically weighted Cohen's kappa (Cohen, 1968), the percentage of exact agreement, and the percentage of classifications differing by no more than one competency level. These analyses were conducted to determine whether Deming regression calibration improved score agreement, reduced systematic measurement differences, and enhanced the consistency of competency classifications between CAT and AC assessments.

RESULTS AND DISCUSSION

The first stage of the series of analysis of the results of this study was to evaluate the initial psychometric characteristics in the raw data (pre-calibration diagnostic) for the entire study population (N = 391). This analysis aimed to map the distribution of the baseline score on an assessment scale of 1 to 5 while testing the strength of the natural correlation between the Computer Adaptive Test (CAT) instrument and the Assessment Center (AC) method before being administered a score equalization intervention. The results of the descriptive statistical test revealed that the mean value of the AC score was in the range of 2.8299 to 2.9086 with a standard deviation (SD) ranging from 0.5397 to 0.5996, while the average raw CAT score was in the range of 2.6968 to 2.7965 with a standard deviation between 0.5400 to 0.6383. In terms of the validity of convergent constructs, the two instruments showed a very strong linear directional relationship and consistency of the sequence of the assessment rating across nine competency aspects. The Pearson linear correlation coefficient (r) was consistently recorded to be above 0.92 (range 0.9216 to 0.9445) and the Spearman rating correlation coefficient (rho) was in the range of 0.9167 to 0.9531, with the highest correlation achieved in the Communication aspect. However, the

evaluation of the absolute value alignment and mean difference confirmed the existence of a negative systematic bias (Bias_CAT_minus_AC) problem ranging from -0.11 to -0.15 points across all aspects of competence. The existence of this negative bias causes the initial concordance coefficient of Lin's CCC (0.8856 to 0.9268) to be held below the Pearson correlation value due to the shift of the data distribution point from the ideal identity line (Y = X), as summarized in full in Table 1 (see Table 1).

Table 1. Descriptive Characteristics, Natural Correlations, and Systematic Bias of Pre-Calibrated Raw Scores (N = 391)

Competency Aspect	N	AC Red	SD AC	CAT Red	Elementary School CAT	Pearson (r)	Spearman (ρ)	Lin's CCC	Bias (CAT–AC)
Integrity	391	2.8645	0.5752	2.7340	0.5912	0.9362	0.9284	0.9120	-0.13
Cooperation	391	2.8971	0.5884	2.7440	0.6134	0.9400	0.9231	0.9097	-0.15
Communication	391	2.8299	0.5996	2.6968	0.6026	0.9445	0.9531	0.9219	-0.13
Result Orientation	391	2.8523	0.5642	2.7006	0.5524	0.9316	0.9287	0.8983	-0.15
Public Service	391	2.8760	0.5486	2.7404	0.5694	0.9318	0.9244	0.9045	-0.13
Development of Others	391	2.9086	0.5990	2.7965	0.6383	0.9439	0.9335	0.9268	-0.11
Managing Changes	391	2.8830	0.5397	2.7290	0.5400	0.9216	0.9167	0.8856	-0.15
Decision Making	391	2.8734	0.5776	2.7330	0.5846	0.9344	0.9215	0.9078	-0.14
Glue of the Nation	391	2.8568	0.5962	2.7445	0.6262	0.9255	0.9337	0.9091	-0.11

The practical impact of a constant negative bias of -0.11 to -0.15 points on the raw continuous score is very evident when the decimal score is converted or rounded into the category of round competency levels (Level 1 to Level 5). The results of the initial categorical agreement evaluation showed that the Quadratic Weighted Kappa value was in the range of 0.6943 to 0.7952 (Substantial Agreement category) with the percentage of absolute agreement ranging from 72.63% in the Cooperation aspect to 82.86% in the Communication aspect. The initial classification security parameters have actually been very well maintained because the Agreement within 1 level index reached a value of 1.0 (100%) in all nine competency aspects, which proves that none of the assessors experienced a difference in prediction of missing more than one level of competence to the actual AC level. However, an examination of the direction of disagreement showed a very sharp disparity between the number of participants who were rated lower by CAT (Lebih_Rendah_CAT) than the number of participants who were rated higher by CAT (Lebih_Tinggi_CAT). As detailed in Table 2 and Table 3 (see Table 2 and Table 3), in the Integrity aspect, there were 57 cases (18 people from AC Level 3 downgraded to CAT Level 2, and 39 people from AC Level 4 downgraded to CAT Level 3) who were disadvantaged due to the stricter nature of the raw CAT instrument, while only 18 participants obtained a higher level. Similar inequality even reached 98 to 9 assessments in the Cooperation aspect, so that six aspects of competence were categorized as Feasible Post-Calibration which is mandatory through the geometric correction of Deming Regression.

Table 2. Evaluation of Raw Categorical Agreement and Competency Level Shift Direction (N = 391)

Competency Aspect	N	Weighted Kappa	Exact Agreement	Within 1 Level	Lower CAT	Higher CAT	Initial Screening Status
Integrity	391	0.7797	80.82%	100.0%	57	18	High Raw Agreement
Cooperation	391	0.6943	72.63%	100.0%	98	9	Feasible Post-Calibration
Communication	391	0.7952	82.86%	100.0%	57	10	High Raw Agreement
Result Orientation	391	0.7355	79.28%	100.0%	68	13	Feasible Post-Calibration
Public Service	391	0.7403	78.26%	100.0%	74	11	Feasible Post-Calibration
Development of Others	391	0.7359	73.66%	100.0%	80	23	Feasible Post-Calibration
Managing Changes	391	0.7023	78.77%	100.0%	74	9	Feasible Post-Calibration
Decision Making	391	0.7080	74.17%	100.0%	86	15	Feasible Post-Calibration
Glue of the Nation	391	0.7769	78.01%	100.0%	71	15	High Raw Agreement

Table 3. Contingency Matrix (Confusion Matrix) Classification of Raw Competency Level on Integrity Aspects (N = 391)

Actual Level of AC \ Raw CAT Prediction	CAT Level 1	CAT Level 2	CAT Level 3	CAT Level 4	CAT Level 5	Total Actual AC
AC Level 1	5	0	0	0	0	5
AC Level 2	0	87	3	0	0	90
AC Level 3	0	18	193	15	0	226
AC Level 4	0	0	39	31	0	70
AC Level 5	0	0	0	0	0	0
Total Raw CAT Prediction	5	105	235	46	0	391 (Diagonal = 316 / 80.82%)

To correct the underestimation bias in the CAT instrument without ignoring the psychometric fact that both the CAT (X) score and the AC (Y) score contain variance in measurement error, score equalization modeling was performed using the Deming Regression method in the training sample (N_train = 261). In contrast to the ordinary least squares regression (Ordinary Least Squares) which only minimizes vertical residues, the Deming Regression minimizes the number of squares of the orthogonal (perpendicular) distance of observation points to the regression line by taking into account the error variance ratio of the two instruments of lambda = 1.0. Mathematically, the estimation of the slope parameters (Slope, beta_1), the fundamental shift constant (Intercept, beta_0), and the calibration score prediction function (AC_pred) are formulated through the following Equations (1), Equations (2), and Equations (3):

$$\beta_1 = [\ (s_yy - \lambda \cdot s_xx) + \sqrt{((s_yy - \lambda \cdot s_xx)^2 + 4 \cdot \lambda \cdot s_xy^2)}\ ] / (2 \cdot s_xy) \quad (1)$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{X} \quad (2)$$

$$AC_pred = \beta_0 + \beta_1 \times CAT \quad (3)$$

In the above formulation, s_xx is the sample variance of the CAT score, s_yy is the sample variance of the AC score, s_xy expresses the sample covariance between the CAT score and the AC score, while $\bar{X}$ and $\bar{Y}$ represent the average value of CAT and AC scores in the training data group, respectively. The results of the estimation of the Deming Regression parameters on the nine competency aspects as presented in Table 4 (see Table 4) show that the value of the constant (Intercept, beta_0) is in a small and stable positive range, which is between 0.07717 (Outcome Orientation) to 0.33937 (Other Person Development). Meanwhile, the regression line slope coefficient (Slope, beta_1) is all very close to the identity value of 1.0, which ranges from 0.92077 to 1.02863. This parameter structure empirically proves that CAT instruments do not experience extreme proportional scale distortions, but rather only require a subtle linear translation shift through a combination of positive Intercept and Slope adjustment to raise the estimated CAT score to be precisely aligned with the AC gold standard. As an operational illustration of the Communication aspect (AC_pred = 0.14350 + 0.99511 × CAT), an assessor who obtains a CAT score of 4.00 will be mapped to an AC prediction score of 4.1239.

Table 4. Intercept, Slope, and Deming Regression Calibration Equations on Training Data (N_train = 261, λ = 1.0)

Competency Aspect	N_train	λ (Lambda)	Intercept (β_0)	Slope (β_1)	Linear Calibration Equation (AC_pred)
Integrity	261	1.0	0.28398	0.94854	AC_pred = 0.28398 + 0.94854 × CAT
Cooperation	261	1.0	0.29863	0.95211	AC_pred = 0.29863 + 0.95211 × CAT
Communication	261	1.0	0.14350	0.99511	AC_pred = 0.14350 + 0.99511 × CAT
Result Orientation	261	1.0	0.07717	1.02863	AC_pred = 0.07717 + 1.02863 × CAT
Public Service	261	1.0	0.26102	0.95778	AC_pred = 0.26102 + 0.95778 × CAT
Development of Others	261	1.0	0.33937	0.92077	AC_pred = 0.33937 + 0.92077 × CAT
Managing Changes	261	1.0	0.19467	0.98813	AC_pred = 0.19467 + 0.98813 × CAT
Decision Making	261	1.0	0.18979	0.98221	AC_pred = 0.18979 + 0.98221 × CAT
Glue of the Nation	261	1.0	0.33208	0.92325	AC_pred = 0.33208 + 0.92325 × CAT

The most crucial test stage in this calibration study was the independent validation evaluation using a test dataset (validation dataset, N_validation = 130) that was not involved in the model parameter formation process at all. The test results as presented in Table 5 and visualized in Figure 1 (see Table 5 and Figure 1) demonstrated a very consistent improvement in psychometric performance across nine competency aspects. The post-calibration Intraclass Correlation

Coefficient (ICC Calibrated) consistency reliability coefficient increased to the range of 0.9241 to 0.9542, which exceeded the threshold of the very strong reliability criterion (> 0.90) (Koo & Li, 2016). The most important achievement is seen in the surge in the post-calibration Lin's CCC concordance coefficient (CCC Calibrated) which now rises to the range of 0.9236 to 0.9538, closely aligned with the ICC value. This increase in Lin's CCC proves geometrically that the Deming Regression transform has successfully shifted the prediction line back above the 45-degree identity line (Y = X). In addition, the continuous prediction error rate is proven to be very small with Mean Absolute Error (MAE) values ranging from 0.1265 to 0.1460 and Root Mean Square Error (RMSE) between 0.1808 to 0.2166. Given that the standard deviation of the actual data is in the range of 0.54 to 0.60, the ratio of RMSE values that are less than one-third of the standard actual deviation (RMSE << SD) confirms the high level of prediction precision and the absence of extreme errors (Predictions beyond 1.5 points = 0). The mean error was successfully reduced to near zero in the range of -0.0325 to 0.0426 with a narrow Bland-Altman Limits of Agreement (LoA) range of around -0.40 to +0.40, as well as the significance value of the equivalence test of Two One-Sided Tests of TOST p = 0.000 (< 0.05) which establishes the status of Equivalent = True in all aspects of competence.

Table 5. Results of Continuous Prediction Accuracy Evaluation and Equivalency Test on Independent Validation Data (N_val = 130)

Competency Aspect	N_val	ICC Calibrated	CCC Calibrated	MAE	RMSE	Mean Error	TOST p	Equivalent	Error > 1.5
Integrity	130	0.9351	0.9346	0.1402	0.2051	-0.0325	0.000	True	0
Cooperation	130	0.9542	0.9538	0.1440	0.1889	0.0426	0.000	True	0
Communication	130	0.9497	0.9494	0.1393	0.1808	-0.0085	0.000	True	0
Result Orientation	130	0.9383	0.9378	0.1382	0.1912	0.0085	0.000	True	0
Public Service	130	0.9286	0.9281	0.1360	0.1996	0.0294	0.000	True	0
Development of Others	130	0.9308	0.9304	0.1460	0.2166	0.0174	0.000	True	0
Managing Changes	130	0.9241	0.9236	0.1437	0.2038	0.0250	0.000	True	0
Decision Making	130	0.9411	0.9406	0.1265	0.1839	0.0023	0.000	True	0
Glue of the Nation	130	0.9403	0.9399	0.1370	0.2037	0.0276	0.000	True	0

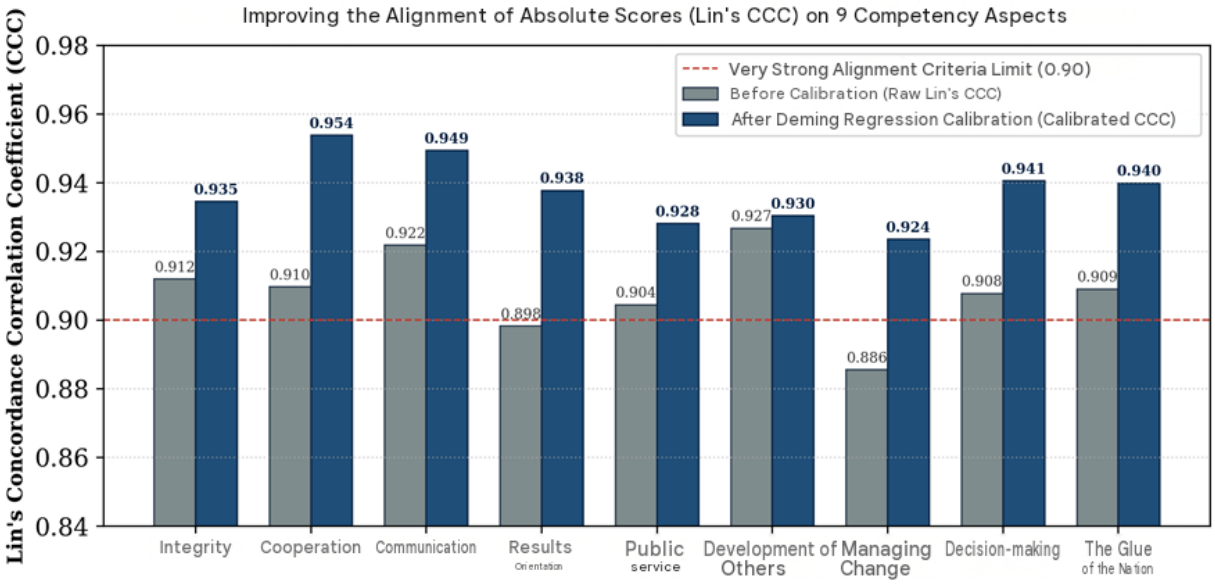

Figure 1. Comparison of Concordance Coefficients (Lin's CCC) Before and After Calibration of Deming Regression on 9 Competency Aspects
(Source: Primary Data Analysis, 2026)

In the perspective of theoretical discussion and comparison of previous studies, the success of Deming Regression in equalizing CAT scores with AC is in line with the psychometric postulates of Kolen and Brennan (2014) and the findings of Linnet (1993) who affirm that the comparison procedure of two behavioral measurement instruments must take into account the variance of measurement errors on both axes of the variable. In contrast to the classical OLS regression approach which often experiences slope attenuation bias due to error-free predictor assumptions, the selection of a balanced error variance ratio (lambda = 1.0) in Deming Regression has been shown to be able to maintain symmetry and equity as required by Angoff (1984). These findings also expand the literature on the use of Computer Adaptive Testing in human resource assessment (van der Linden & Glas, 2010; Kantrowitz et al., 2018), where the difference in modality between computerized situational responses and interpersonal behavioral observations in the Assessment Center often results in differences in the severity of the score, but still measures an equivalent latent construct. By correcting the negative bias of -0.11 to -0.15 points to a near-zero residue (-0.03 to +0.04), organizations can avoid dozens of candidates from the risk of unfair decline in competency levels. The main strength of this study lies in the integrity of the data intact (N = 391 with no data lost), the strict separation of independent validation samples, and the use of comprehensive multi-metric continuous and categorical evaluation. The limitation of this study is that modeling was carried out on one population of the organization assuming a constant error variance ratio (lambda = 1.0) and the unavailability of observations at the highest extremes of the category (Level 5), so that testing of population invariance across agencies and periodic monitoring using post-implementation Bland-Altman graphs is still recommended.

CONCLUSION

Based on the diagnostic analysis, calibration modeling, and independent validation results, this study concluded that the computer adaptive testing (CAT) instrument implemented through the Adaptive Public Works Test (ATPU) demonstrated strong convergent evidence and high consistency in competency rankings relative to the assessment center (AC) reference method. However, the raw CAT scores exhibited systematic negative bias, indicating the need for linear calibration to improve score comparability. Before calibration, CAT scores were consistently 0.11–0.15 points lower than their corresponding AC scores, potentially resulting in the underclassification of participants' competency levels. Deming regression, with an assumed measurement error variance ratio of $\lambda = 1.0$, was applied to the training sample ($n = 261$), yielding nine competency-specific linear calibration equations with slope coefficients ranging from 0.92077 to 1.02863 and intercepts ranging from 0.07717 to 0.33937. Independent validation ($n = 130$) demonstrated strong absolute agreement between calibrated CAT scores and AC reference scores, with intraclass correlation coefficients (ICCs) and Lin's concordance correlation coefficients (CCCs) ranging from 0.92 to 0.95. Prediction errors remained relatively low, with mean absolute errors (MAEs) of 0.12–0.14 points and root mean squared errors (RMSEs) of 0.18–0.21 points. No individual prediction errors exceeded 1.5 points, and the two one-sided tests (TOST) indicated statistical equivalence within the prespecified equivalence margins ($p < 0.05$). These findings suggested that the nine Deming regression calibration equations provided a promising approach to aligning CAT scores with the AC reference scale, supporting more consistent competency assessment and classification in large-scale personnel evaluations. Nevertheless, operational implementation should be supported by continued monitoring of calibration stability, classification agreement, and measurement fairness across different participant groups and assessment periods.

FUNDING

The research and preparation of this article was carried out independently without funding support from external parties.

CONFLICT OF INTEREST

The author states that there are no conflicts of interest, whether personal, commercial, or financial, that can affect the implementation and results of this study.

AUTHOR'S CONTRIBUTION

Rud Nokeo[1]*, Fieghie Fadila[2] dan Gayatri Indah Puspitorini[3] contributed to research design, data collection and processing, system development, result analysis, and article manuscript preparation.

GENERATIVE ARTIFICIAL INTELLIGENCE USAGE STATEMENT

In the process of compiling this article, the author uses generative artificial intelligence (AI) as a tool in language refinement, writing structure, and grammar improvement. The use of AI

is carried out on a limited basis and remains under the supervision of the author. All data, analysis, interpretation, and scientific substance in this article have been checked and verified by the author. The author is solely responsible for the accuracy, authenticity, and final content of the manuscript.

REFERENCES

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125–153. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Boyatzis, R. E. (2008). Competencies in the 21st century. *Journal of Management Development, 27*(1), 5–12. https://doi.org/10.1108/02621710810840730

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Deming, W. E. (1943). *Statistical adjustment of data*. John Wiley & Sons.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2), i–41. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

González, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-51824-4

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*(4), 405–411. https://doi.org/10.1111/j.1468-2389.2007.00399.x

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management, 41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Kantrowitz, T. M., Tuzinski, K. A., & Raines, J. M. (2018). *Global assessment trends report 2018*. SHL.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science & Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lin, L. I. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268. https://doi.org/10.2307/2532051

Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. *Clinical Chemistry, 39*(3), 424–432. https://doi.org/10.1093/clinchem/39.3.424

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using catR and mstR packages*. Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Passing, H., & Bablok, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. *Journal of Clinical Chemistry and Clinical Biochemistry, 21*(11), 709–720. https://doi.org/10.1515/cclm.1983.21.11.709

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657–680. https://doi.org/10.1007/BF01068419

Spencer, L. M., & Spencer, S. M. (1993). *Competence at work: Models for superior performance*. John Wiley & Sons.

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Lawrence Erlbaum Associates.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-85461-8

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1–27. https://doi.org/10.2458/v2i1.12351

Williams, K. N., Lazzara, E. H., Sadighi, M., Chandran, N., Joshi, K., Raj, S., Shields, I., Nichols, B., Testa, D., & Hernandez, J. (2024). Integrating behavioral assessment in instructional design for competency-based medical education. *Frontiers in Medicine*, *11*, 1432319.