https://doi.org/10.7251/SSH2602226K **UDC:** 37.091.3:797.21-053.6
Original scientific paper ***Originalni naučni rad***

Comparing Dichotomous and Polytomous Instruments in CDM-Based Swimming Technique Assessment

Fifukha Dwi Khory, Ali Maksum, Suroto, Setiyo Hartoto, Taufiq Hidayat, Bayu Budi Prakoso
Universitas Negeri Surabaya, Surabaya, East Java, Indonesia

Correspondence:
Fifukha Dwi Khory
Universitas Negeri Surabaya, Indonesia, fifukhakhory@unesa.ac.id

Abstract: *Assessing complex motor skills across physical performance domains poses distinct methodological challenges for applying the Cognitive Diagnostic Model (CDM), particularly regarding the response format of observational instruments. This study compared inter-rater reliability, model fit, diagnostic specificity, and criterion validity between dichotomous and polytomous instruments for assessing front crawl technique. A total of 280 physical education students from Universitas Negeri Surabaya participated. The dichotomous instrument (10 items, three attributes: Arm Action, Body & Leg, Breath & Synchronization) was adapted from the Crawl Technique Observation Sheet and analyzed using GDINA. The polytomous instrument (20 items, four dimensions) was analyzed using pGDINA with a three-attribute Q-matrix, after Body Position and Leg Action were mapped into a single Body & Leg attribute based on biomechanical justification, structural equivalence, and sample adequacy. The dichotomous instrument showed higher inter-rater reliability (Fleiss' κ = 0.663) than the polytomous instrument (ICC = 0.538), with a Classification Accuracy of 0.924 and higher validity against swimming speed (ρ = 0.507 vs. 0.288). The polytomous instrument produced higher diagnostic entropy (2.362 vs. 1.568 bits) and a more even latent profile distribution (largest profile 40.7% vs. 62.5%), indicating better diagnostic specificity. Cross-instrument profile consistency was low (Cohen's κ = 0.370), with 31.4% of borderline students classified differently by both instruments. A trade-off emerged: the dichotomous instrument excelled in reliability, classification accuracy, and criterion validity, whereas the polytomous instrument excelled in diagnostic depth, confirming that CDM applications in motor skill assessment face challenges distinct from cognitive education contexts.*

Keywords: *Cognitive Diagnostic Model; GDINA; polytomous instrument; inter-rater reliability; front crawl swimming*

Introduction

Physical education has long assessed motor skills using quantitative outcomes such as speed, distance, or holistic scores. These measures capture how well a student performs but say little about why. A student who misses a target swimming time might struggle with arm coordination, body position, or breathing, yet a single total score cannot distinguish among these issues (Collares, 2022; de la Torre & Akbay, 2019; Popham, 2003; Rupp et al., 2010). This gap between performance and diagnosis has led researchers to more sophisticated measurement approaches. (de la Torre, 2011) GDINA model, part of the broader family of Cognitive Diagnostic Models (CDM), offers one such approach. Rather than placing students on a single ability continuum as Item Response Theory does, CDM classifies them according to a profile of mastered and unmastered latent attributes (Rupp et al., 2010). A swimmer might master Arm Action but not Breathing, and that distinction is precisely what makes CDM profiles useful for instructors designing targeted remediation (Collares, 2022; de la Torre & Akbay, 2019).

CDM has proven effective in cognitive academic domains such as reading, mathematics, and language (Ma & de la Torre, 2016; Meng et al., 2023), but its application to psychomotor skills remains largely untested. Front crawl swimming, which requires the simultaneous coordination of arm and leg movements, body rotation, and breathing (Maglischo, 2003), is a natural candidate because its components can be modeled as distinct yet interacting latent attributes (de la Torre, 2011; Rupp et al., 2010). GDINA accommodates this without assuming a particular interaction structure in advance (Ma & de la Torre, 2020), and its polytomous extension, pGDINA, links graded item categories to attributes via a categorical Q-matrix (Ma & de la Torre, 2016). However, forcing polytomous responses into a dichotomous mold consistently discards diagnostic information (Chen & de la Torre, 2018).

The harder problem is practical: how should qualitative, observation-based components of the technique be scored so that CDM can actually use them? Dichotomous instruments score each component as present or absent, which maps cleanly onto standard CDM (de la Torre, 2011) and underlies the Crawl Technique Observation Sheet

developed by (de Medeiros Vidal et al., 2021). By contrast, polytomous instruments use graded ordinal scales, which should, in principle, capture finer gradations in movement quality (Chen & de la Torre, 2018; Ma & de la Torre, 2016). That richness comes at a cost: ordinal judgments are harder to make consistently than binary ones (Moskal & Leydens, 2000), and Sampaio Peres et al. (2023) found that standardized observation criteria alone were insufficient to achieve adequate reliability in assessing swimming technique. Whether the theoretical advantage of polytomous instruments survives this reliability cost, and which format performs better when consistency, classification accuracy, and predictive validity are weighed together, remain open empirical questions.

This study addresses four questions: do the two instruments yield consistent attribute profiles; which yields more diagnostically specific information; how much information is lost through dichotomization; and which demonstrates stronger criterion validity relative to swimming speed and Distance Per Stroke. The findings aim to inform CDM-based assessment methodology in physical education and to offer practical guidance for researchers extending CDM into motor skill domains that remain largely unexplored in the psychometric literature.

Method

Ethical Approval and Participants

This study received ethical approval under Certificate No. 005/C8/G/I/2025 from the Research and Community Service Unit of STKIP PGRI Bangkalan, dated January 5, 2025. All participants provided written informed consent before participating. A total of 280 active physical education students from Universitas Negeri Surabaya participated, comprising 202 men (72.1%) and 78 women (27.9%), with a mean age of 19.4 ± 0.9 years, drawn from 18 course sections. All participants had completed coursework in basic swimming and/or aquatic instruction, and each swam 50 meters of front crawl at a self-selected, comfortable pace, following the procedure described by de Medeiros Vidal et al. (2021).

Instruments

Two instruments were used in this study. The dichotomous instrument was adapted from the Crawl Technique Observation Sheet (de Medeiros Vidal et al., 2021) and comprised 10 binary-scored items, reclassified into three latent attributes based on biomechanical relevance: Arm Action, Body & Leg, and Breath & Synchronization. Content validity of this reclassification was confirmed by seven experts using Aiken's V coefficient (Aiken, 1985), with all items deemed valid (V = 0.857 to 1.000). The polytomous instrument, developed specifically for this study, consists of 20 items spanning four dimensions of front crawl technique: Body Position, Leg Action, Arm Action, and Breathing. Each item was scored using a two-stage procedure: raters first assigned a broad category (Poor, Fair, or Good), then refined their judgment within that category to arrive at a final score from one to five (Moskal & Leydens, 2000).

Table 1. *Q-Matrix of the Dichotomous Instrument*

Item	Description (original wording)	A1 (AA)	A2 (BL)	A3 (BS)
1a	Hand entrance position in relation to head	1	0	0
1b	Elbow height in relation to hand in stroke recovery	1	0	0
1d	Elbow extension at end of stroke	1	0	0
1e	Elbow height in relation to hand during the stroke's propulsive phase	1	0	0
1c	Hand entrance location in relation to the body's central line	0	1	0
2a	Leg amplitude	0	1	0
3a	Horizontal alignment (body position)	0	1	0
4a	Head movement and arm coordination while breathing	0	0	1
5a	Body rolling	0	0	1
5b	Arm coordination	0	0	1

Note: *AA = Arm Action; BL = Body & Leg; BS = Breath & Synchronization.*

Although the polytomous instrument was designed and retained with four dimensions (Body Position, Leg Action, Arm Action, Breathing), the Q-matrix used for CDM analysis was planned around a three-attribute structure,

with the Body Position and Leg Action dimensions mapped into a single combined Body & Leg attribute pending data support. This methodological decision, made at the analysis stage rather than the instrument design stage, rested on three criteria. First, biomechanical relevance: the horizontal body position in front-crawl swimming is actively maintained by the kicking action as a stabilizer, so the two dimensions are conceptually closely linked (Maglischo, 2003). Second, structural equivalence: this mapping yields three attributes aligned with the dichotomous instrument, enabling a fair comparison between the two. Third, sample adequacy: four attributes would generate sixteen potential latent classes, requiring a much larger sample for reliable estimation than the eight latent classes produced by three attributes (Gudicha et al., 2016; Sorrel et al., 2023).

Data Collection and Performance Measurement

Data collection for the dichotomous instrument took place directly at a standard 50-meter pool, with each participant videotaped from start to finish and rated simultaneously by three trained raters; the final score for each item was determined by the mode of the three ratings (Stemler, 2004). The same videos were then scored for the polytomous instrument by ten trained raters organized into five fixed pairs, with the final score taken as the mean of each pair's two ratings, rounded to the nearest integer (Ma & de la Torre, 2016; Stevens, 1946; Zumbo & Kroc, 2019). Dichotomous raters completed three training sessions involving joint scoring of pilot videos, criterion discussion, and perceptual calibration, continuing until $\kappa \geq 0.61$ agreement was reached before actual scoring began. Polytomous raters underwent six calibration cycles using Kendall's W, with agreement rising consistently from $W = 0.085$ in the first cycle to $W = 0.538$ in the sixth, as a prerequisite before scoring the main videos.

Swimming speed and Distance Per Stroke, two valid and reliable indicators for assessing criterion validity (Jürimäe et al., 2007), were derived from the same videos. Elapsed time was measured with a manual stopwatch from start to finish over the 50-meter distance, yielding $v = 50$ m / elapsed time. Stroke count used a separate-arm counting method, in which each right- or left-arm pull was counted as a separate stroke, so one complete cycle corresponded to two raw counts; Distance Per Stroke was calculated as DPS = 50 m / number of cycles, which is algebraically equivalent to DPS = 100 m / raw stroke count, as reported in the Data Analysis section. Speed reflects overall performance as the result of all technique components acting simultaneously, whereas DPS measures propulsive efficiency per cycle and is more sensitive to specific aspects of technique quality (Maglischo, 2003).

Data Analysis

Inter-rater reliability for the dichotomous instrument was computed using Fleiss' Kappa, and for the polytomous instrument, the Intraclass Correlation Coefficient was calculated with a two-way mixed, absolute agreement model. Both were computed using the irr package in R (Gamer et al., 2019). The GDINA model for the dichotomous instrument was estimated using the GDINA() function from the GDINA package (Ma & de la Torre, 2020), and the pGDINA model for the polytomous instrument was estimated using the gdina() function from the CDM package (George et al., 2016), with a 20-by-3 Q-matrix. Model fit was evaluated using M_2, RMSEA, AIC, and BIC statistics, and classification accuracy was computed using the CA() function. Cross-instrument comparison was conducted at four levels: diagnostic specificity, assessed by the number of unique profiles and latent profile entropy; inter-attribute correlation, assessed using Spearman correlations for the dichotomous instrument and polychoric correlations for the polytomous instrument; profile consistency, assessed using Cohen's Kappa between the two instruments' attribute classifications (Cohen, 1960); and criterion validity, assessed using Spearman correlations between profile scores and both swimming speed and Distance Per Stroke.

Results

Descriptive Statistics

Participants' mean swimming speed was 0.810 ± 0.222 m/s (range 0.269 to 1.736 m/s), with a mean elapsed time of 66.94 ± 20.86 seconds. Mean Distance Per Stroke was 1.805 ± 0.395 m/cycle. The substantial variation in both indicators reflects a diverse range of skill levels among participants, consistent with the sample needed for diagnostic analysis.

Inter-Rater Reliability: Dichotomous and Polytomous Instruments

The dichotomous instrument showed high inter-rater reliability, with a mean Fleiss' $\kappa = 0.663$ (Substantial) across three raters. All 10 items reached the substantial category ($\kappa \geq 0.60$), except item 5a, which fell in the moderate range ($\kappa = 0.597$). The highest coefficient was observed for item 1e ($\kappa = 0.758$, Arm Action). By contrast, the polytomous instrument showed lower reliability, with a mean ICC = 0.538 (Moderate) across two raters. The Body Position dimension showed the widest variation (ICC ranging from 0.414 to 0.629), with several items falling into the poor category, while the Leg Action dimension reached the highest ICC (0.592). This substantial reliability gap between the two instruments ($\kappa = 0.663$ vs. ICC = 0.538) reflects an inherent difficulty in maintaining consistent ordinal judgments for complex motor skills.

Cognitive Diagnostic Model — Dichotomous Instrument (GDINA)

The methodological decision to map Body Position and Leg Action into a single Body & Leg attribute on the polytomous Q-matrix was empirically supported: the polychoric correlation between the two dimensions was 0.945, the highest among all attribute pairs. With N = 280 and three attributes yielding eight latent classes, a sample-to-class ratio of 10.4 was considered adequate for stable estimation.

The GDINA model achieved very high Classification Accuracy at both the pattern level (0.924) and the attribute level (mean 0.968). The latent profile distribution showed a dominant "masters none of the attributes" profile (0-0-0) at 62.5%, followed by the "masters all" profile (1-1-1) at 20.7%. Attribute mastery proportions followed the order Arm Action (36.4%) > Body & Leg (31.1%) > Breath & Synchronization (22.9%), indicating that breathing coordination was the most difficult component to master. The M_2 statistic could not be stably estimated due to zero cell frequencies in the bivariate item matrix, a known technical limitation for highly asymmetric response distributions (Maydeu-Olivares & Joe, 2006); therefore, model fit evaluation relied on RMSEA and individual item fit, both of which indicated adequate fit (Lei & Li, 2016).

Table 2. *GDINA Model Fit and Attribute Profile Distribution (Dichotomous)*

Indicator	Value	Interpretation
AIC	2,356.64	—
BIC	2,454.78	—
Pattern CA	0.924	Excellent
Attribute CA (mean)	0.968	Excellent
CA — Arm Action	0.970	Excellent
CA — Body & Leg	0.975	Excellent
CA — Breath & Sync	0.961	Excellent
Dominant profile: 0-0-0 (masters none)	62.5% (n=175)	—
Profile 1-1-1 (masters all)	20.7% (n=58)	—
Arm Action mastery proportion	36.4%	—
Body & Leg mastery proportion	31.1%	—
Breath & Sync mastery proportion	22.9%	Lowest

Cognitive Diagnostic Model — Polytomous Instrument (pGDINA)

The pGDINA model yielded a mean RMSEA of 0.079, right at the threshold for an adequate fit, with the Arm Action (0.083) and Breathing (0.084) dimensions slightly exceeding it. The latent profile distribution was more evenly distributed than for the dichotomous instrument, with the dominant 0-0-0 profile accounting for only 40.7% and the 1-1-1 profile for 27.9%. Polychoric correlations among attributes were very high (0.886–0.975), indicating substantial multicollinearity.

Table 3. *pGDINA Model Fit and Attribute Profile Distribution (Polytomous)*

Indicator	Value	Interpretation
Deviance	8,973.043	—
AIC	9,147.04	—
BIC	9,463.27	—
Mean RMSEA	0.079	Adequate (< 0.08)
RMSEA — Body & Leg	0.074	Good
RMSEA — Arm Action	0.083	Borderline
RMSEA — Breathing	0.084	Borderline
Dominant profile: 0-0-0	40.7% (n=114)	—
Profile 1-1-1 (masters all)	27.9% (n=78)	—
Body & Leg mastery	44.3%	—
Arm Action mastery	45.7%	—
Breathing mastery	40.0%	Lowest

Diagnostic Specificity and Information Loss from Dichotomization

The polytomous instrument detected all eight possible latent profiles (2^3), whereas the dichotomous instrument detected only six of the eight, as profiles 0-1-0 and 0-1-1 did not appear in the data. Profile distribution entropy was substantially lower for the dichotomous instrument (1.568 bits) than for the polytomous instrument (2.362 bits, a 50.6% difference), indicating that the dichotomous distribution was far more concentrated on a single dominant profile. Mean inter-attribute correlation was much lower for the dichotomous instrument (r = 0.751) than for the polytomous instrument (r = 0.928), indicating that dichotomous attributes were more discriminable from one another. A total of 88 students (31.4%) were identified as having borderline profiles on the polytomous instrument, and their attribute classification consistency between the two instruments ranged only from 50.0% to 64.8%, confirming that dichotomization discards significant information, particularly for students at intermediate ability levels.

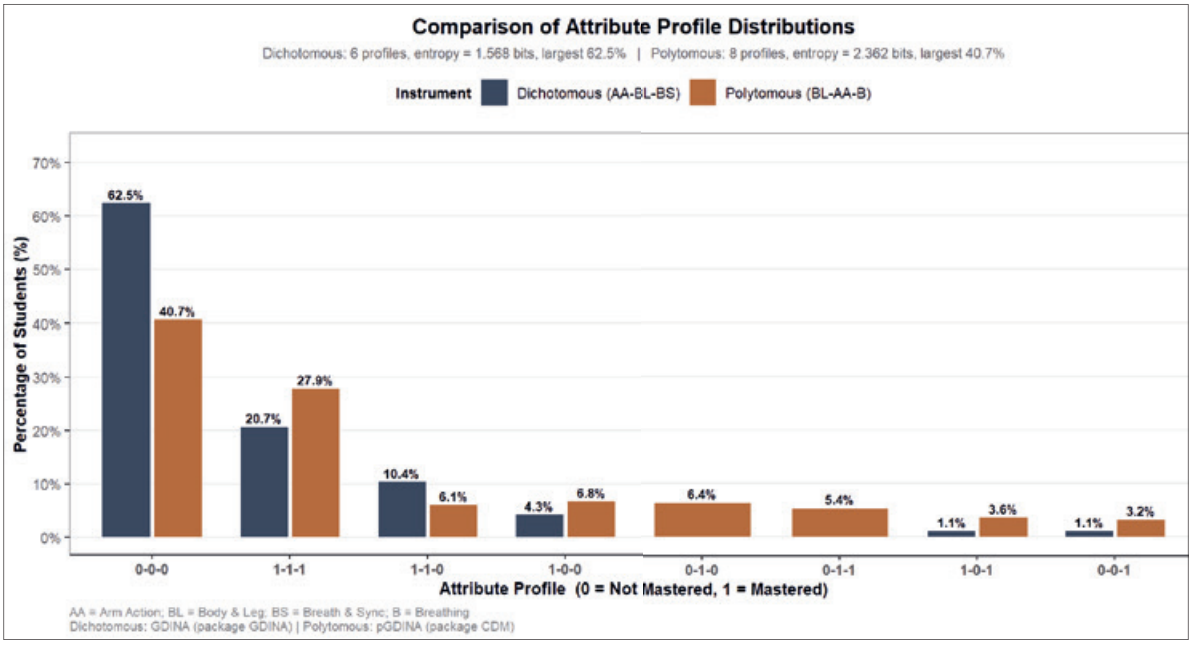

Figure 1. *Comparison of attribute profile distributions from GDINA (dichotomous instrument) and pGDINA (polytomous instrument) analyses. AA = Arm Action; BL = Body & Leg; BS = Breath & Synchronization; B = Breathing. Dichotomous: 6 profiles detected, entropy = 1.568 bits, largest profile 62.5%. Polytomous: 8 profiles detected, entropy = 2.362 bits, largest profile 40.7%.*

Comparative Diagnostic Performance Between Instruments

Table 4. *Summary of Diagnostic Comparison*

Indicator	Dichotomous	Polytomous
Inter-rater reliability	κ = 0.663 (Substantial)	ICC = 0.538 (Moderate)
Model fit (RMSEA)	Good*	0.079 (Adequate)
Classification Accuracy (pattern)	0.924	Not available†
Unique profiles detected	6 of 8	8 of 8
Diagnostic entropy	1.568 bits	2.362 bits (+50.6%)
Dominant profile (largest)	62.5%	40.7%
Inter-attribute correlation (mean)	r = 0.751	r = 0.928
Cross-instrument κ consistency	0.370 (Fair)	-
Validity vs. v (ρ)	0.507 (p < 0.001)	0.288 (p < 0.001)
Validity vs. DPS (ρ)	0.265 (p < 0.001)	0.255 (p < 0.001)
Borderline students	-	88 (31.4%)
Borderline classification consistency	54.5–64.8%	-

**M_2 could not be stably estimated due to zero cell frequencies in the bivariate item matrix, resulting from a highly uneven response distribution; fit evaluation relied on RMSEA and available item fit statistics. †The CDM package does not provide a CA function equivalent for pGDINA.*

Inter-Attribute Correlation

Mean inter-attribute correlation was much lower for the dichotomous instrument (r = 0.751) than for the polytomous instrument (r = 0.928). The lower correlation for the dichotomous instrument indicates that its attributes were more discriminable from one another, a more favorable psychometric property for diagnostic purposes. The very high polychoric correlations among polytomous attributes (0.886–0.975) indicate substantial multicollinearity, which may limit the distinctiveness of diagnostic profiles.

External Criterion Validity

The dichotomous instrument showed higher criterion validity for swimming speed (ρ = 0.507) than the polytomous instrument (ρ = 0.288), although both were statistically significant (p < 0.001). The gap was smaller for Distance Per Stroke (ρ = 0.265 vs. 0.255). At the individual attribute level, Breath & Sync (dichotomous) correlated most strongly with DPS (ρ = 0.225), whereas Breathing (polytomous) correlated most strongly with both v (ρ = 0.278) and DPS (ρ = 0.245), confirming that breathing coordination was the technique component contributing most to swimming efficiency.

Table 5. *Spearman Correlations with External Criteria*

Attribute/Score	Dichotomous (ρ vs. v)	Dichotomous (ρ vs. DPS)	Polytomous (ρ vs. v)	Polytomous (ρ vs. DPS)
Total profile score	0.507***	0.265***	0.288***	0.255***
Arm Action	0.475***	0.261***	0.239***	0.194***
Body & Leg	0.470***	0.216***	0.239***	0.197***
Breath & Sync / Breathing	0.431***	0.225***	0.278***	0.245***

**** p < 0.001. v = swimming speed (m/s); DPS = Distance Per Stroke (m/cycle).*

Discussion

Diagnostic Trade-off Between Reliability and Specificity

The central finding of this study confirms a fundamental trade-off between dichotomous and polytomous instruments within a CDM framework for motor skill assessment. The dichotomous instrument excelled in reliability (κ = 0.663 vs. ICC = 0.538), classification accuracy (CA = 0.924), and criterion validity (ρ = 0.507 vs. 0.288 against swimming speed). By contrast, the polytomous instrument produced a more diagnostically informative profile distribution, with higher entropy (2.362 vs. 1.568 bits), more unique profiles (8 vs. 6), and a smaller dominant profile (40.7% vs. 62.5%). This trade-off has a clear theoretical basis: binary judgments are easier to make consistently because they involve only two decisions, whereas a five-point ordinal scale requires raters to draw finer distinctions, thereby increasing ambiguity and reducing agreement (Moskal & Leydens, 2000). Although the two-stage scoring procedure was designed to mitigate this, a mean ICC of 0.538 still indicates suboptimal consistency, consistent with Sampaio Peres et al. (2023) finding that standardized criteria alone are not sufficient to achieve high reliability in swimming technique assessment.

The strikingly high 0-0-0 dominant profile on the dichotomous instrument (62.5%) warrants cautious interpretation. Beyond reflecting genuinely modest skill levels among participants (consistent with a mean swimming speed of 0.810 m/s), another possibility is that the rigid binary mastery threshold misclassifies intermediate-level students as non-masters, a phenomenon known in the CDM literature as the masking effect of dichotomization (Chen & de la Torre, 2018; Ma & de la Torre, 2016). The borderline analysis supports this possibility: 31.4% of students showed intermediate profiles on the polytomous instrument, yet the dichotomous instrument failed to capture these nuances, with classification consistency for this group ranging from 50.0% to 64.8%. The implication is that the dichotomous instrument risks categorizing intermediate-level students as complete non-masters, even though they may have partially mastered certain components of the technique.

Psychometric Challenges: Multicollinearity and Criterion Validity

The very high polychoric correlations among polytomous attributes (0.886–0.975) indicate substantial multicollinearity, a psychometric challenge because CDM models struggle to distinguish each attribute's unique contribution when attributes are highly correlated, potentially yielding unstable parameter estimates and overlapping latent profiles (de la Torre, 2011). The lower inter-attribute correlation for the dichotomous instrument (r = 0.751) indicates more discriminable attributes, which is a favorable property for CDM analysis. The higher correlations observed for the polytomous instrument likely stem from the close biomechanical interdependence among front crawl technique components (Maglischo, 2003), as well as a possible halo effect in rater judgments that were not fully isolated across dimensions.

The dichotomous instrument's higher criterion validity relative to swimming speed (ρ = 0.507 vs. 0.288) is informative, since swimming speed reflects the simultaneous action of all technique components. A stronger correlation with v indicates that the attribute profile more closely reflects overall technique quality (Jürimäe et al., 2007; Maglischo, 2003). However, this higher validity may partly reflect the dichotomous instrument's more bipolar distribution, whereas the polytomous instrument captures finer gradations in movement quality that correlate less directly with overall speed (Chen & de la Torre, 2018). The merely fair cross-instrument profile consistency (Cohen's κ = 0.370) confirms that the two instruments produce substantively different diagnostic pictures, rather than simply differing in measurement scale (Landis & Koch, 1977).

Implications for Applying CDM to Motor Skill Assessment

This study has two main implications for extending CDM beyond cognitive-academic domains. First, reliability challenges in observation-based motor skill assessment exceed those typical of written cognitive testing because observational judgments are temporal and subjective, making them more susceptible to inter-rater inconsistency. This implies that substantial investment in rater training is a non-negotiable prerequisite before applying CDM to motor domains. Second, the choice of response format should be guided by the intended use: when classification accuracy and predictive validity relative to performance are the priority, the dichotomous instrument with standard GDINA performs better; when the goal is to provide more nuanced feedback on gradations in movement quality, the polytomous instrument offers richer information despite its lower reliability. The low cross-instrument profile consistency

($\kappa = 0.370$) also carries an important methodological warning: researchers should not assume that two instrument formats measure identical constructs simply because they share the same attribute labels. As a technical limitation, the M_2 statistic for the GDINA model could not be stably estimated due to zero cell frequencies in the bivariate item matrix, a known constraint for highly asymmetric response distributions (Maydeu-Olivares & Joe, 2006). This makes the exploration of alternative fit indices, such as bootstrap-based fit statistics, an important direction for future research.

Conclusion

This study confirms a genuine methodological trade-off between dichotomous and polytomous instruments when applying CDM to front crawl swimming assessment. The dichotomous instrument excelled in inter-rater reliability ($\kappa = 0.663$, Substantial), classification accuracy (Pattern CA = 0.924), and criterion validity against swimming speed ($\rho = 0.507$), whereas the polytomous instrument excelled in diagnostic specificity, with higher profile distribution entropy (2.362 vs. 1.568 bits) and more unique profiles detected (8 vs. 6). The merely fair cross-instrument profile consistency (Cohen's $\kappa = 0.370$), together with the finding that 31.4% of borderline-profile students were classified differently by the two instruments, confirms that response format selection is not merely a technical decision but a substantive one that fundamentally shapes diagnostic interpretation. Future research should explore rater training strategies to improve the reliability of polytomous instruments, develop hybrid approaches that integrate the strengths of both instrument types, and replicate these findings across different samples and motor skill domains. This study's primary contribution is providing empirical evidence, grounded in real rather than simulated data, on the application of CDM to motor skill assessment, a domain still rarely explored in the CDM literature, while enriching understanding of the boundaries and potential of CDM as a diagnostic assessment paradigm in physical education.

References

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Chen, J., & de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/FPSYG.2018.01474

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collares, C. F. (2022). Cognitive diagnostic modelling in healthcare professions education: an eye-opener. *Advances in Health Sciences Education*, *27*(2), 427–440. https://doi.org/10.1007/S10459-022-10093-Y

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/S11336-011-9207-7

de la Torre, J., & Akbay, L. (2019). Implementation of cognitive diagnosis modeling using the GDINA r package. *Eurasian Journal of Educational Research*, 171–192. https://eric.ed.gov/?id=EJ1211628

de Medeiros Vidal, J., Tucher, G., Nogueira, L., Novaes, R. C., de Souza Vale, R. G., de Castro, M. O. R., da Silva Medeiros, N. G., & de Cassio Costa Telles, S. (2021). Crawl technique observation sheet for beginning swimmers: an evaluation proposal for swimming teachers. *Motriz: Revista de Educação Física*, *27*, e1021016920. https://doi.org/10.1590/S1980-65742021016920

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement*. CRAN. https://cran.r-project.org/package=irr

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

Gudicha, D. W., Tekle, F. B., & Vermunt, J. K. (2016). Power and Sample Size Computation for Wald Tests in Latent Class Models. *Journal of Classification*, *33*(1), 30–51. https://doi.org/10.1007/S00357-016-9199-1

Jürimäe, J., Haljaste, K., Cicchella, A., Lätt, E., Purge, P., Leppik, A., & Jürimäe, T. (2007). Analysis of Swimming Performance from Physical, Physiological, and Biomechanical Parameters in Young Swimmers. *Pediatric Exercise Science*, *19*(1), 70–81. https://doi.org/10.1123/PES.19.1.70

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lei, P. W., & Li, H. (2016). Performance of Fit Indices in Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied Psychological Measurement*, *40*(6), 405–417. https://doi.org/10.1177/0146621616647954

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *The British Journal of Mathematical and Statistical Psychology*, *69*(3), 253–275. https://doi.org/10.1111/BMSP.12070

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Maglischo, E. W. (2003). *Swimming fastest*. Human Kinetics.

Maydeu-Olivares, A., & Joe, H. (2006). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/S11336-005-1295-9

Meng, Y., Wang, Y., & Zhao, N. (2023). Cognitive diagnostic assessment of EFL learners' listening barriers through incorrect responses. *Fron-*

tiers in Psychology, *14*, 1126106. https://doi.org/10.3389/FPSYG.2023.1126106
Moskal, B. M., & Leydens, J. A. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research, and Evaluation*, *7*(1). https://doi.org/10.7275/Q7RM-GG74
Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. ASCD.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
Sampaio Peres, M. V., Nogueira, F., Ramos, W., Cardoso, T., Nogueira, L., & Tucher, G. (2023). Effect of standardization of observation and evaluation criteria of the front crawl arm stroke technique on inter- and intra-rater reliability. *Journal of Physical Education*, *34*(1). https://doi.org/10.4025/jphyseduc.v34i1.3416
Sorrel, M. A., Escudero, S., Nájera, P., Kreitchmann, R. S., & Vázquez-Lira, R. (2023). Exploring Approaches for Estimating Parameters in Cognitive Diagnosis Models with Small Sample Sizes. *Psych*, *5*(2), 336–349. https://doi.org/10.3390/PSYCH5020023
Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research, and Evaluation*, *9*(1). https://doi.org/10.7275/96JP-XZ07
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/SCIENCE.103.2684.677
Zumbo, B. D., & Kroc, E. (2019). A Measurement Is a Choice and Stevens' Scales of Measurement Do Not Help Make It: A Response to Chalmers. *Educational and Psychological Measurement*, *79*(6), 1184–1197. https://doi.org/10.1177/0013164419844305

Primljen: 24. jul 2026. / Received: July 24, 2026
Prihvaćen: 27. avgust 2026. / Accepted: August 27, 2026

This work is licensed under a **Creative Commons Attribution-NonCommercial 4.0 International License**.