psychiatry international

MDPI

Article

Developmental Sensitivity and Selected Psychometric Properties of the Koppitz Scoring System for the Bender–Gestalt Test: A Preliminary Study in School-Aged Children from the Rabat Region, Morocco

Hicham Lafhal [1,2,3,*], Ahmed Omar Tohami Ahami [1], Siham Goutou [4] and Atmane Rochdi [2]

[1] Laboratory of Biology and Health, Department of Biology, Faculty of Science, Ibn Tofail University, Kenitra 14000, Morocco; ahami_40@yahoo.fr
[2] Laboratory of Natural Resources and Sustainable Development, Department of Biology, Faculty of Science, Ibn Tofail University, Kenitra 14000, Morocco; atmane.rochdi@uit.ac.ma
[3] Psychomotor, Rehabilitation, Higher Institute of Nursing Professions and Health Techniques, Rabat 10000, Morocco
[4] Continuing Education, Children's Hospital, Ibn Sina University Hospital Center, Rabat 10000, Morocco; siham_goutou@hotmail.com
* Correspondence: hicham.lafhal@uit.ac.ma; Tel.: +212-662588874

Abstract

This study examined visuomotor performance and selected psychometric characteristics of the Bender–Gestalt Test using the Koppitz developmental scoring system in Moroccan children aged 7 to 11 years. A cross-sectional study was conducted with 200 children enrolled in public primary schools in the Rabat region, Morocco. Performance was assessed using the Koppitz developmental scoring system. Statistical analyses included descriptive statistics, factorial analyses of variance, post hoc comparisons, inter-rater reliability, and test–retest reliability. The results revealed significant age-related differences in total error scores, with older age groups obtaining lower mean error scores than younger age groups, $F(4, 190) = 29.965$, $p < 0.001$, $\eta p^2 = 0.387$. Designs 3 and 7 produced the highest mean error scores, which may reflect their greater visuospatial and perceptual-motor complexity. Inter-rater reliability demonstrated good agreement (ICC = 0.805, 95% CI [0.669, 0.887], $p < 0.001$), while test–retest reliability demonstrated excellent temporal stability (single-measure ICC(A, 1) = 0.913, 95% CI [0.772, 0.960], Pearson's r = 0.928, $p < 0.001$). These findings support the reproducibility and temporal stability of the Koppitz developmental scoring system. Overall, the findings support its use for the assessment of visuomotor performance in Moroccan school-aged children from the Rabat region and provide preliminary regional reference data that may contribute to the development of future national normative standards.

Keywords: Bender–Gestalt Test; Koppitz scoring system; developmental assessment; psychometric properties; visuomotor integration; Moroccan children; neuropsychological assessment; developmental sensitivity

Check for updates

Academic Editor: Stefano Barlati

Received: 26 May 2026
Revised: 21 July 2026
Accepted: 23 July 2026
Published: 1 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Visuomotor integration is a fundamental neurodevelopmental function involved in many academic skills, including writing, reading, and spatial organization. It reflects the coordinated interaction between visual perception and motor execution, enabling children

to accurately reproduce, organize, and integrate visual stimuli. Difficulties with visuomotor integration have frequently been associated with learning disabilities, particularly during the early school years [1,2]. Thus, the assessment of visuomotor abilities plays an important role in developmental neuropsychology and psychoeducational evaluation.

Among the instruments commonly used to assess visuomotor functioning is the Bender–Gestalt Test. Initially developed by Lauretta Bender in 1938 [3] based on the principles of Gestalt psychology, this test assesses an individual's ability to reproduce geometric designs, thereby providing information on perceptual organization, motor coordination, and developmental maturity. One of the most influential scoring systems associated with the test is the developmental approach proposed by Elizabeth Munsterberg Koppitz [4,5]. This standardized method is based on the identification of several types of errors, including distortions, rotations, integration difficulties, and perseverations. Koppitz's system has been widely used in child neuropsychology for its developmental sensitivity and relevance to clinical and psychoeducational assessment [6].

However, the use of Koppitz's scoring system in different cultural and educational contexts raises several methodological questions. The psychometric properties of an assessment tool may vary depending on the sociocultural, linguistic, and educational characteristics of the populations studied, highlighting the need for preliminary regional reference values [7,8]. Furthermore, most normative and psychometric studies on the Bender–Gestalt Test have been conducted with Western populations, potentially limiting the generalizability of the findings to other cultural contexts [9].

In Morocco, locally validated neuropsychological tools for children remain relatively limited, particularly for visuomotor assessment in school settings. Differences in educational systems, socioeconomic conditions, and early graphic experiences may influence the performance observed among Moroccan children [10]. Therefore, the present study aimed to examine the developmental sensitivity and selected psychometric characteristics of the Bender–Gestalt Test using Koppitz's scoring system in Moroccan children aged 7 to 11 years. More specifically, the study investigated age- and sex-related differences, test–retest reliability, and the distribution of errors across Bender–Gestalt designs.

This study contributes to developmental neuropsychology by providing preliminary regional reference data and additional evidence regarding the developmental sensitivity and reliability of the Koppitz developmental scoring system in Moroccan school-aged children. It also offers methodological and cross-cultural contributions while providing clinically relevant information to support the assessment of visuomotor development in Moroccan school-aged children.

2. Materials and Methods

2.1. Participants and Sampling

This study adopted a quantitative, cross-sectional, descriptive design to examine age-related differences and selected psychometric characteristics of the Bender–Gestalt Test, using the Koppitz developmental scoring system, among Moroccan school-aged children in the Rabat region. Participants were recruited using a non-probability purposive sampling method. The final sample consisted of 200 Moroccan children, including 92 boys and 108 girls, aged 7 to 11 years, enrolled in four public primary schools in the Rabat region. Schools were selected in consultation with educational authorities to obtain a relatively homogeneous socioeconomic sample.

Children were eligible for inclusion if they were enrolled in mainstream education and had no known diagnosis of a neurodevelopmental disorder or specific learning disorder, according to school records and teacher reports. Children were excluded if they had uncorrected visual impairment, a known neurological disorder, or a diagnosed intellectual

disability or if they had previously been referred for psychological assessment or special education services.

Participants were divided into five age groups, corresponding to ages 7, 8, 9, 10, and 11 years, with 40 children in each age group. Age groups were defined according to chronological age at the time of assessment. Thus, the 7-year group included children aged 7.00–7.99 years, the 8-year group included children aged 8.00–8.99 years, and so forth up to 11.99 years. The distribution of participants by age and sex is presented in Table 1.

Table 1. Distribution of participants by age and sex.

Age	Boys, *n*	Girls, *n*	Total, *n*
7 years	20	20	40
8 years	17	23	40
9 years	19	21	40
10 years	18	22	40
11 years	18	22	40
Total	92	108	200

Note. Values are presented as frequencies. Boys represented 46% of the sample, and girls represented 54%.

2.2. Recruitment and Ethical Procedures

The research protocol was reviewed and approved by the Ethics Board of the Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, Morocco, under reference number C64/20. The study was conducted with prior authorization from the relevant educational authorities and school administrations. Participation was voluntary, and no financial or academic incentives were offered. Written informed consent was obtained from parents or legal guardians before any assessment was conducted. Parents were informed about the aims and scope of the study, the confidentiality and anonymity of the collected data, and their right to refuse their child's participation or withdraw from the study at any time. Children also provided verbal assent before participating in the assessment.

Assessments were conducted by six trained examiners, including four psychomotor therapists and two licensed psychologists, all of whom had prior experience in neuropsychological assessment in school settings. To reduce examiner-related bias and ensure standardized administration conditions, assessments were conducted in pairs in each school.

Before data collection, the examiners completed an intensive three-day training program led by an experienced neuropsychologist. The training included a theoretical review of the Bender–Gestalt Test and the Koppitz scoring system, practical administration workshops using role-play, and collective scoring sessions based on test protocols. Specialized examiners rather than school teachers administered the assessments because standardized administration and scoring require specific competencies in neuropsychological assessment and application of the Koppitz scoring criteria.

Data collection took place between January 2026 and May 2026.

2.3. Test Administration

The instrument used was the Bender–Gestalt Test, consisting of nine 10 × 15 cm cards displaying geometric designs. The standard designs corresponding to cards A, 1, 2, 3, 4, 5, 6, 7, and 8 were used. Each card contained a visual design that was presented successively to the participants, who were asked to reproduce it on a blank A4 sheet of paper.

The test was administered individually in a quiet classroom according to a standardized protocol established during examiner training. Administration conditions were designed to create a supportive testing environment, reduce sources of distraction, maintain the child's motivation, and allow short breaks if signs of fatigue appeared.

Instructions were given in Moroccan Arabic, in accordance with the administration procedures described by Koppitz [5]. Simple age-appropriate explanations were provided to ensure that the child understood the task, without influencing the reproduction of the designs.

The cards were presented vertically and sequentially, with each card introduced after the previous one was completed. Participants were allowed to handle the cards, but they were instructed not to rotate them. No time limit was imposed during test administration.

2.4. Scoring Procedure

The researcher scored test protocols according to the criteria of the Koppitz developmental scoring system [5]. This system is based on the identification of 30 specific error items, coded according to their presence or absence in children's graphic productions. These errors are grouped into four main categories: distortion, rotation, integration difficulties, and perseveration. Each developmental error item is scored dichotomously (0 = absent, 1 = present). Because each Bender–Gestalt design comprises a different number of developmental error items, the possible score range varies across designs, whereas the total developmental error score ranges from 0 to 30. For each protocol, a total error score was calculated by summing all observed errors. A higher score indicated a greater number of visuomotor errors and therefore lower visuomotor performance. When ambiguity arose during scoring, the original Koppitz manual was consulted to ensure consistent application of the scoring criteria. The distribution of these error items across Bender–Gestalt designs and error categories is presented in Table 2.

Table 2. Distribution of Koppitz developmental error items by Bender–Gestalt design, error category, and possible score range.

Design	Form Distortion	Rotation	Integration Difficulty	Perseveration	Possible Score (Min–Max)
A	1a and 1b	2	3		0–4
1	4	5		6	0–3
2		7	8	9	0–3
3	10	11	12a and 12b		0–4
4		13	14		0–2
5	15	16	17a and 17b		0–4
6	18a and 18b		19	20	0–4
7	21a and 21b	22	23		0–4
8	24	25			0–2
Total error items per category	10	8	9	3	
Total developmental error score			30		0–30

Note. Error categories are based on the Koppitz developmental scoring system [5].

2.5. Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 26.0 (IBM Corp., Armonk, NY, USA; RRID:SCR_019096). Descriptive statistics were computed for all study variables. Before conducting factorial analyses of variance (ANOVA), the assumptions of homogeneity of variances and normality of residuals were evaluated. Homogeneity of variances was assessed using Levene's test. Normality of the residuals for the factorial ANOVA model of the total Bender–Gestalt error score was evaluated using the Kolmogorov–Smirnov and Shapiro–Wilk tests, together with visual inspection of histograms and normal Q–Q plots. For the design-specific ANOVAs, graphical diagnostics (histograms and normal Q–Q plots) were examined to verify the adequacy of the residual distributions. Because the study design was balanced, with equal sample sizes across age groups (n = 40), minor departures from normality would not have substantially affected the robustness of the ANOVA analyses. Given the balanced study design and equal sample sizes across age

groups (n = 40), ANOVA was considered sufficiently robust for analysing the bounded developmental error scores.

Factorial ANOVAs were conducted to examine the effects of age and sex on total Bender–Gestalt developmental error scores and on each of the nine Bender–Gestalt designs. When the main effect of age was statistically significant, Bonferroni-adjusted pairwise comparisons were performed to identify differences between age groups. Because separate factorial ANOVAs were conducted for each Bender–Gestalt design, the Holm–Bonferroni procedure was applied separately to the sets of age, sex, and age × sex interaction effects to control the family-wise error rate. Statistical significance was set at $p < 0.05$.

Test–retest reliability was evaluated in a stratified random subsample of 50 children (10 from each age group) who completed the Bender–Gestalt Test on two occasions separated by a three-week interval. The subsample was selected to ensure representation of all age groups and both sexes and was comparable to the full study sample with respect to age-group distribution and sex. Agreement between test and retest scores was assessed using a single-measure intraclass correlation coefficient (ICC) based on a two-way mixed-effects model with absolute agreement (ICC(A, 1)). Pearson's correlation coefficient was additionally reported to facilitate comparison with previous studies. The mean paired difference between test and retest scores, the 95% confidence interval (CI), and the standard error of measurement (SEM) were also calculated.

To evaluate inter-rater reliability, a separate stratified random subsample of 50 protocols (10 from each age group) was independently rescored by a second trained examiner who was blinded to the original ratings. This subsample was independent of the one used for the test–retest reliability analysis. Agreement for the total Koppitz developmental error score was assessed using a single-measure intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement (ICC(A, 1)).

3. Results

To evaluate the reproducibility of the Koppitz scoring procedure, inter-rater reliability was assessed in a stratified random subsample of 50 protocols (10 children from each age group). The protocols were independently rescored by a second trained examiner blinded to the original ratings. Agreement for the total Koppitz developmental error score was evaluated using the intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement.

The analysis demonstrated good inter-rater reliability (single-measure ICC(A, 1) = 0.805, 95% CI [0.669, 0.887], $p < 0.001$), indicating good agreement between the two independent raters and supporting the reproducibility of the scoring procedure.

Test–retest reliability was evaluated in a subsample of 50 children aged 7 to 11 years reassessed after a three-week interval. Agreement between test and retest scores was assessed using the intraclass correlation coefficient (ICC) based on a two-way mixed-effects model with absolute agreement. Pearson's correlation coefficient was also calculated to facilitate comparison with previous studies.

Test–retest reliability demonstrated excellent temporal stability (single-measure ICC(A, 1) = 0.913, 95% CI [0.772, 0.960], Pearson's r = 0.928, $p < 0.001$). Pearson's correlation coefficient (r = 0.928, $p < 0.001$) was additionally reported to facilitate comparison with previous studies using the Koppitz scoring system.

The mean total error score decreased slightly from the initial assessment ($M = 3.34$, $SD = 1.69$) to the retest session ($M = 3.02$, $SD = 1.39$), corresponding to a mean paired difference of 0.32 points (95% CI [0.13, 0.51], $t(49) = 3.47$, $p = 0.001$). This small reduction in error scores may reflect a modest practice effect associated with repeated administration while maintaining overall score stability.

The standard measurement error (SEM) was estimated at 0.55, providing an estimate of the measurement error associated with the total Koppitz developmental error score.

Table 3 summarizes the test–retest reliability of the Bender–Gestalt Test using the Koppitz scoring system.

Table 3. Test–retest reliability of the Bender–Gestalt Test using the Koppitz scoring system over a three-week interval.

Session	*N*	Mean (*M*)	Standard Deviation (*SD*)	Pearson's *r*	*p*	Single-Measure ICC(A, 1)
Test	50	3.34	1.686	0.928	<0.001	0.913 (95% CI [0.772, 0.960])
Retest	50	3.02	1.392			

Mean paired difference = 0.32 (95% CI [0.13, 0.51]); SEM = 0.55. Note. ICC(A, 1) = single-measure intraclass correlation coefficient based on a two-way mixed-effects model with absolute agreement; SEM = standard error of measurement.

Descriptive statistics for the total Bender–Gestalt error score according to age and sex are presented in Table 4.

Table 4. Descriptive statistics for the total Bender–Gestalt error score by age and sex.

Age	Girls *M* (*SD*)	Boys *M* (*SD*)	Total *M* (*SD*)
7 years	5.25 (1.86)	5.15 (1.76)	5.20 (1.79)
8 years	4.04 (1.22)	2.94 (1.03)	3.57 (1.26)
9 years	3.71 (1.06)	2.42 (1.12)	3.10 (1.26)
10 years	2.86 (1.61)	2.39 (1.46)	2.65 (1.55)
11 years	2.05 (1.29)	1.94 (1.21)	2.00 (1.24)
Total	3.56 (1.77)	3.01 (1.77)	3.31 (1.79)

Note. *M* = mean; *SD* = standard deviation.

The results showed significant age-related differences in total error scores, with older age groups obtaining lower mean error scores than younger age groups. The mean score decreased from *M* = 5.20 (*SD* = 1.79) at age 7 to *M* = 2.00 (*SD* = 1.24) at age 11, indicating lower mean error scores among the older age groups. At the descriptive level, girls obtained slightly higher mean total error scores than boys (*M* = 3.56, *SD* = 1.77 vs. *M* = 3.01, *SD* = 1.77). These descriptive differences were most apparent at ages 8 and 9 but became less pronounced in the older age groups. The age-group differences in total error scores are illustrated in Figure 1.

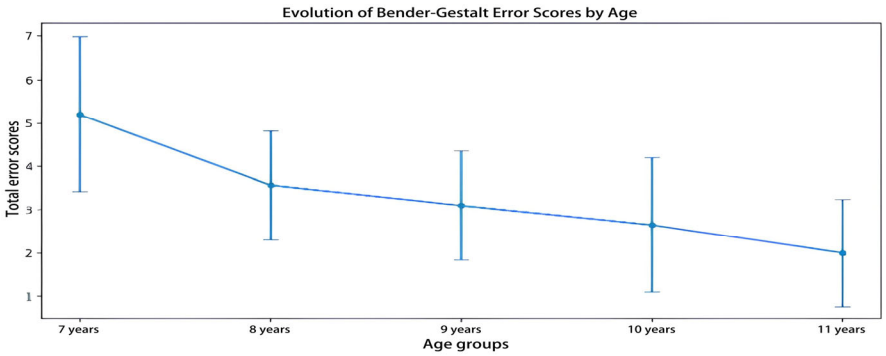

Figure 1. Mean total Bender–Gestalt error scores by age group. Note. Error bars represent standard deviations.

To examine the effects of age and sex on Bender–Gestalt performance, a factorial analysis of variance was conducted after verifying the assumptions of homogeneity of variances and normality of residuals. Visual inspection of histograms and normal Q–Q

plots, together with the Kolmogorov–Smirnov test (D = 0.060, $p = 0.075$) and the Shapiro–Wilk test (W = 0.989, $p = 0.150$), indicated no significant deviations from normality for the residuals of the factorial ANOVA model of the total Bender–Gestalt error score. For the design-specific ANOVA models, graphical inspection of the residual distributions also revealed no substantial departures from normality. Homogeneity of variances was confirmed using Levene's test (F(9, 190) = 1.673, $p = 0.098$). These findings supported the use of factorial ANOVA for both the total-score and design-specific analyses. The results are presented in Table 5.

Table 5. Levene's test and factorial analysis of variance results for total Bender–Gestalt error scores by age and sex.

Analysis	*df*	F	*p*	ηp^2
Levene's test	(9, 190)	1.673	0.098	—
Age	(4, 190)	29.965	<0.001	0.387
Sex	(1, 190)	9.603	0.002	0.048
Age × Sex	(4, 190)	1.589	0.179	0.032
Overall model	(9, 190)	15.097	<0.001	0.417

Note. *df* = degrees of freedom; F = F statistic; *p* = significance level; ηp^2 = partial eta squared. The non-significant Levene's test indicates that the assumption of homogeneity of variances was met. Partial eta-squared values represent the effect sizes of age, sex, and the age × sex interaction on total Bender–Gestalt error scores.

The factorial ANOVA revealed a significant main effect of age, F(4, 190) = 29.965, $p < 0.001$, $\eta p^2 = 0.387$, indicating significant age-related differences in total Bender–Gestalt error scores, with older age groups obtaining lower mean error scores than younger age groups. A statistically significant main effect of sex was also observed, F(1, 190) = 9.603, $p = 0.002$, $\eta p^2 = 0.048$. However, the effect size was small, indicating that sex accounted for only a limited proportion of the variance in total error scores. The age × sex interaction was not significant, F(4, 190) = 1.589, $p = 0.179$, $\eta p^2 = 0.032$, suggesting that the pattern of age-group differences was comparable for boys and girls. The overall model was significant, F(9, 190) = 15.097, $p < 0.001$, $\eta p^2 = 0.417$, accounting for 41.7% of the variance in total Bender–Gestalt error scores.

Bonferroni-adjusted pairwise comparisons (Table 6) showed that children aged 7 years obtained significantly higher total error scores than all older age groups (8, 9, 10, and 11 years; all adjusted $p < 0.001$). Children aged 8 years also had significantly higher error scores than children aged 10 years (adjusted $p = 0.034$) and 11 years (adjusted $p < 0.001$), whereas the difference between the 8- and 9-year groups was not statistically significant (adjusted $p = 1.000$). Similarly, children aged 9 years differed significantly from those aged 11 years (adjusted $p = 0.005$) but not from those aged 10 years (adjusted $p = 1.000$). No significant difference was observed between the 10- and 11-year groups (adjusted $p = 0.386$), suggesting a progressive reduction in visuomotor errors with age that tended to stabilize between 10 and 11 years.

Mean errors for each Bender–Gestalt design were analyzed across age groups to examine age-related differences in visuomotor performance. As shown in Table 7, mean errors generally decreased with increasing age. Designs 3 and 7 showed the highest mean error scores across age groups, suggesting that these designs were the most difficult for the children. In contrast, designs 1, 4, and 6 showed relatively low mean error scores, particularly among older children. Seven-year-old children obtained the highest error scores on most designs, especially design 3 ($M = 0.95$, $SD = 0.639$), design 7 ($M = 0.92$, $SD = 0.526$), and design 8 ($M = 0.85$, $SD = 0.483$). Children aged 8 to 11 years generally obtained lower and more stable error scores than the youngest age group. The lowest error levels were observed among 11-year-old children, particularly for design 6 ($M = 0.03$,

SD = 0.158) and design 2 (*M* = 0.08, *SD* = 0.267). These findings indicate age-related differences in design-specific Bender–Gestalt error scores.

Table 6. Bonferroni-adjusted pairwise comparisons of total Bender–Gestalt error scores across age groups.

Comparison	Mean Difference (I − J)	95% CI	Adjusted *p*
7 vs. 8	1.63	[0.74, 2.51]	<0.001
7 vs. 9	2.10	[1.21, 2.99]	<0.001
7 vs. 10	2.55	[1.66, 3.44]	<0.001
7 vs. 11	3.20	[2.31, 4.09]	<0.001
8 vs. 9	0.47	[−0.41, 1.36]	1.000
8 vs. 10	0.92	[0.04, 1.81]	0.034
8 vs. 11	1.57	[0.69, 2.46]	<0.001
9 vs. 10	0.45	[−0.44, 1.34]	1.000
9 vs. 11	1.10	[0.21, 1.99]	0.005
10 vs. 11	0.65	[−0.24, 1.54]	0.386

Note. Mean differences are presented as (I − J), with positive values indicating higher error scores for the first age group. Pairwise comparisons were adjusted using the Bonferroni correction. CI = confidence interval. Significant comparisons (adjusted $p < 0.05$) indicate statistically significant differences between age groups.

Table 7. Mean errors by Bender–Gestalt design across age groups.

Design	7 Years *M* (*SD*)	8 Years *M* (*SD*)	9 Years *M* (*SD*)	10 Years *M* (*SD*)	11 Years *M* (*SD*)	Total *M* (*SD*)
Design A	0.23 (0.423)	0.27 (0.452)	0.27 (0.452)	0.23 (0.423)	0.15 (0.362)	0.23 (0.422)
Design 1	0.45 (0.552)	0.15 (0.362)	0.10 (0.304)	0.08 (0.267)	0.13 (0.335)	0.18 (0.398)
Design 2	0.50 (0.506)	0.40 (0.496)	0.33 (0.474)	0.27 (0.452)	0.08 (0.267)	0.31 (0.466)
Design 3	0.95 (0.639)	0.68 (0.474)	0.48 (0.506)	0.30 (0.464)	0.62 (0.490)	0.61 (0.557)
Design 4	0.27 (0.452)	0.18 (0.385)	0.20 (0.405)	0.23 (0.423)	0.13 (0.335)	0.20 (0.401)
Design 5	0.63 (0.667)	0.63 (0.490)	0.50 (0.506)	0.35 (0.483)	0.18 (0.385)	0.46 (0.538)
Design 6	0.40 (0.496)	0.18 (0.385)	0.15 (0.362)	0.22 (0.423)	0.03 (0.158)	0.19 (0.397)
Design 7	0.92 (0.526)	0.87 (0.648)	0.85 (0.622)	0.85 (0.622)	0.50 (0.506)	0.80 (0.602)
Design 8	0.85 (0.483)	0.23 (0.423)	0.22 (0.423)	0.15 (0.362)	0.18 (0.385)	0.33 (0.490)

Note. *M* = mean; *SD* = standard deviation. Each age group included $n = 40$ children, for a total sample of $N = 200$. Values represent mean errors observed for each Bender–Gestalt design.

Factorial analyses of variance (age × sex) were conducted separately for each of the nine Bender–Gestalt designs to identify the designs most sensitive to age-related variation. As shown in Table 8, the effect of age differed across designs. Significant age effects were observed for designs 1, 2, 3, 5, 6, 7, and 8, indicating greater sensitivity to age-related differences. The largest age-related effect sizes were found for design 8 ($\eta p^2 = 0.300$), design 3 ($\eta p^2 = 0.158$), and design 1 ($\eta p^2 = 0.122$), suggesting that these designs were particularly sensitive to age-related differences in visuomotor and perceptual-motor performance. In contrast, designs A and 4 did not show significant age effects, indicating lower sensitivity to age-related differences within the studied age range. Before correction for multiple testing, small sex effects were observed for designs 2 and 6 ($p = 0.047$ and $p = 0.048$, respectively). However, these effects did not remain statistically significant after Holm–Bonferroni correction. No significant age × sex interaction was found for any design (all $p > 0.05$), indicating similar patterns of age-group differences in boys and girls.

Table 8. Factorial analysis of variance results for the nine Bender–Gestalt designs by age and sex.

Design	Age Effect *F*	*p*	ηp^2	Sex Effect *F*	*p*	ηp^2	Age × Sex Effect *F*	*p*	ηp^2
Design A	0.447	0.775	0.009	1.951	0.164	0.010	1.041	0.388	0.021
Design 1	6.571	<0.001	0.122	0.000	0.988	0.000	0.928	0.449	0.019
Design 2	4.987	<0.001	0.095	4.014	0.047	0.021	0.821	0.513	0.017
Design 3	8.888	<0.001	0.158	1.992	0.160	0.010	0.801	0.526	0.017
Design 4	0.794	0.530	0.016	0.318	0.573	0.002	1.201	0.312	0.025
Design 5	5.732	<0.001	0.108	0.701	0.403	0.004	1.200	0.312	0.025
Design 6	5.192	<0.001	0.099	3.972	0.048	0.020	0.623	0.646	0.013
Design 7	3.308	0.012	0.065	0.200	0.655	0.001	0.449	0.773	0.009
Design 8	20.322	<0.001	0.300	1.037	0.310	0.005	1.110	0.353	0.023

Note. *F* = F statistic; *p* = significance level; ηp^2 = partial eta squared. Separate factorial ANOVAs were conducted for each Bender–Gestalt design. The Holm–Bonferroni procedure was applied separately to the age, sex, and age × sex interaction effects to control the family-wise error rate. After correction, only the age effects for designs 1, 2, 3, 5, 6, 7, and 8 remained statistically significant. No design-specific sex effect or age × sex interaction remained statistically significant after correction.

4. Discussion

The findings of the present study are broadly consistent with observations reported in several international studies using the Koppitz scoring system for the Bender–Gestalt Test. As reported in the original developmental work of Elizabeth Munsterberg Koppitz [5], lower error scores were observed in the older age groups, consistent with expected age-related differences in visuomotor performance. The mean scores obtained in the present Moroccan sample also appear relatively close to those reported in studies conducted in South Africa [11], Hong Kong [12], and Italy [13]. However, some cross-cultural differences remain observable, particularly when compared with Iranian data [14] and some South African findings [11], where slightly higher error levels have sometimes been reported.

These findings are consistent with the hypothesis that older age groups demonstrate better visuomotor integration and visuoconstructive performance, in line with expected neuropsychological maturation. The lower error scores observed in the older age groups may reflect age-related differences in visuomotor coordination, perceptual organization, fine motor control, and attentional abilities involved in graphic reproduction. In this sense, the significant age effect observed in the present study confirms the developmental sensitivity of the Bender–Gestalt Test scored according to the Koppitz system for assessing visuomotor performance in school-aged children.

The analysis of errors by design provided additional information beyond the total score. Designs showing the highest error rates, particularly designs 3 and 7, may reflect greater visuospatial and perceptual-motor complexity, although these underlying cognitive processes were not directly assessed in the present study. This finding is consistent with previous observations concerning the relative difficulty of certain geometric configurations of the test [5,13]. Moreover, the design-specific analyses showed that some designs were more sensitive to age-related variation than others. This supports the clinical usefulness and developmental sensitivity of examining not only the total error score but also the qualitative distribution of errors across individual Bender–Gestalt designs.

Regarding sex differences, a statistically significant effect was observed for the total developmental error score; however, the associated effect size was small ($\eta p^2 = 0.048$), indicating limited practical significance. Girls obtained slightly higher mean error scores than boys in the present sample, although this difference accounted for only a small proportion of the total variance. Therefore, this finding should be interpreted with caution. It may reflect sample-specific characteristics or educational and sociocultural influences rather than inherent biological differences between boys and girls. Moreover, previous studies using the Bender–Gestalt Test have reported inconsistent findings regarding sex

differences, with several studies reporting minimal or no meaningful differences. In the present study, age emerged as a substantially stronger determinant of visuomotor performance than sex.

The test–retest findings further support the reliability of the Koppitz developmental scoring system in the Moroccan context. In addition to the excellent temporal stability demonstrated by Pearson's correlation coefficient, the present study showed excellent agreement between test and retest scores based on the intraclass correlation coefficient, low measurement error as reflected by the standard error of measurement, and good inter-rater reliability. Together, these findings support the stability, precision, and reproducibility of the total Koppitz developmental error score in Moroccan school-aged children.

The Koppitz developmental scoring system is composed of heterogeneous visuomotor error indicators rather than items designed to measure a single latent construct. Consequently, agreement-based indices such as inter-rater reliability and test–retest reliability are more appropriate indicators of reliability for this type of developmental assessment.

The cross-cultural differences observed across studies should be interpreted cautiously. Although variations in educational practices, sociocultural context, familiarity with graphic activities, cognitive stimulation, and methodological differences between studies may contribute to the cross-cultural differences observed, these variables were not directly assessed in the present study. Therefore, these explanations should be regarded as hypotheses rather than conclusions supported by the current findings. Future studies should investigate these factors directly to better understand their contribution to visuomotor performance in Moroccan children.

From a clinical and educational perspective, the present findings provide preliminary support for the use of the Bender–Gestalt Test scored according to the Koppitz system as part of a comprehensive neuropsychological assessment of visuomotor development in Moroccan school-aged children. However, the present study did not evaluate convergent, discriminant, predictive, or diagnostic validity, nor did it include clinical populations. Consequently, the clinical utility of the test should be considered preliminary. The Bender–Gestalt Test should not be used as a standalone diagnostic or screening instrument but rather as one component of a broader multidisciplinary assessment incorporating clinical observation, developmental history, educational information, and complementary standardized measures.

Several limitations should be acknowledged. First, the study used a non-probability purposive sampling method and included children from four public schools in the Rabat region, which may limit the generalizability of the findings to all Moroccan children. The sample did not include children from private schools, rural areas, or diverse socioeconomic backgrounds. Although the present findings provide useful preliminary regional reference data, future multicenter studies using nationally representative samples will be necessary before national normative standards can be established.

Second, the developmental reference sample included only typically developing children. Consequently, the preliminary regional reference values reported in the present study should not be generalized directly to clinical populations but rather considered as developmental reference values for comparison with future clinical investigations.

Third, although the study examined developmental sensitivity and test–retest reliability, it did not assess convergent, discriminant, predictive, or diagnostic validity, nor did it include clinical populations. Consequently, the present findings do not allow conclusions regarding the diagnostic accuracy or screening performance of the Bender–Gestalt Test. Accordingly, broader psychometric conclusions should be drawn with caution until these complementary forms of validity have been established.

Fourth, potentially relevant contextual variables, such as educational practices, socioeconomic characteristics, familiarity with graphic activities, and environmental stimulation, were not directly assessed. Therefore, the explanations proposed for the observed cross-cultural differences should be interpreted as plausible hypotheses rather than conclusions supported by the present data. Future studies should directly examine these contextual factors to better understand their potential contribution to visuomotor performance in Moroccan children.

Another limitation concerns the analyses of design-specific error scores. Because the scores for individual Bender–Gestalt designs are discrete and bounded, with relatively restricted score ranges that vary according to the number of developmental error items, parametric ANOVA may be less optimal for these outcomes than for the total developmental error score. Although the balanced study design and assumption diagnostics supported the use of ANOVA, generalized linear models may provide a more appropriate analytical framework for design-specific outcomes in future studies. Therefore, the findings related to individual Bender–Gestalt designs should be interpreted as exploratory.

Finally, an additional limitation concerns the assessment of test–retest reliability. Although the same standardized scoring criteria were applied across both administrations, the retest protocols were not scored under blinded conditions. Consequently, scorer expectancy bias cannot be completely excluded. Future studies should incorporate blind scoring procedures.

Future studies should replicate the present inter-rater reliability findings in larger and independent samples. They should also examine convergent, discriminant, predictive, and diagnostic validity by comparing Bender–Gestalt performance with other standardized measures of visuomotor integration, executive functioning, academic achievement, and neurodevelopmental status. Studies including children with neurodevelopmental disorders, learning disabilities, and other clinical conditions will be essential to determine the diagnostic accuracy and clinical utility of the Koppitz scoring system.

Future studies should also directly examine the influence of educational, socioeconomic, and sociocultural factors on visuomotor performance to clarify their role in explaining cross-cultural variability in Bender–Gestalt performance.

Taken together, these findings support the developmental sensitivity and reliability of the Koppitz scoring system in Moroccan school-aged children from the Rabat region while highlighting the need for broader normative and validation studies before national reference standards can be established.

For descriptive comparison with previous international studies, the published developmental error scores obtained using the Koppitz developmental scoring system are summarized in Table 9.

Table 9. Descriptive comparison of published developmental error scores on the Bender–Gestalt Test across international studies using the Koppitz developmental scoring system.

Study	Country	Age Range	N	Youngest Group ($M \pm SD$)	Oldest Group ($M \pm SD$)	Developmental Trend
Koppitz [5]	USA	5–10	523	6.45 ± 3.12	1.30 ± 1.55	↓
Ghassemzadeh [14]	Iran	5–11	101	12.93 (SD NR)	2.14 (SD NR)	↓
Viljoen et al. [11]	South Africa	6–11	882	4.70 (SD NR)	2.01 (SD NR)	↓
Mazzeschi & Lis [13]	Italy	3.6–11.5	1065	13.56 ± 3.74	2.35 ± 1.98	↓
Chan [12]	Hong Kong	4.5–8.5	748	8.77 ± 3.20	2.01 ± 1.42	↓
Present study	Morocco	7–11	200	5.20 ± 1.79	2.00 ± 1.24	↓

Note. Comparisons are descriptive only. The included studies differed in sampling procedures, participant characteristics, age ranges, study periods, and cultural contexts. Consequently, these data should not be interpreted as evidence of cross-cultural measurement equivalence or cross-cultural applicability of the Koppitz scoring system. The downward arrow (↓) indicates a decrease in developmental error scores with increasing age. NR = standard deviation not reported in the original publication.

5. Conclusions

This study examined age-related differences and selected psychometric characteristics of the Bender–Gestalt Test scored according to the Koppitz developmental scoring system in Moroccan school-aged children from the Rabat region. The findings suggest significant age-related differences in visuomotor performance, with lower error scores observed in the older age groups, together with good inter-rater reliability and excellent test–retest reliability, supporting the reliability and developmental sensitivity of the Koppitz scoring system in this population. Although a statistically significant sex effect was observed, its practical importance was limited compared with the substantially stronger effect associated with age.

The findings are broadly consistent with previous international studies while emphasizing the importance of developing locally adapted normative references for Moroccan children. The present study provides preliminary regional reference data and evidence supporting the developmental sensitivity of the Koppitz scoring system among Moroccan school-aged children from the Rabat region. However, these findings should be interpreted as preliminary and require confirmation through larger nationally representative studies before national normative standards can be established.

Future research should include nationally representative samples, children from different educational and socioeconomic backgrounds, and clinical populations. Additional studies examining convergent, discriminant, predictive, and diagnostic validity, together with comparisons with other standardized measures of visuomotor and neuropsychological functioning, will further strengthen the evidence supporting the clinical and educational use of the Bender–Gestalt Test in Morocco.

Author Contributions: Conceptualization, H.L. and A.R.; methodology, H.L., S.G. and A.R.; validation, S.G. and A.R.; formal analysis, H.L.; investigation, H.L. and A.O.T.A.; data curation, H.L. and A.O.T.A.; writing—original draft preparation, H.L.; writing—review and editing, H.L., A.O.T.A., S.G. and A.R.; supervision, S.G. and A.R.; project administration, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Board of the Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, Morocco (protocol code: No. C64/20; approval date: 24 February 2021).

Informed Consent Statement: Written informed consent was obtained from the parents or legal guardians of all participants involved in the study. Verbal assents were also obtained from the children before assessment.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available due to confidentiality and ethical restrictions involving child participants but are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the participating children, parents, school staff, and educational authorities for their cooperation and support during the study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANOVA Analysis of Variance

References

1. Sortor, J.M.; Kulp, M.T. Are the results of the Beery-Buktenica Developmental Test of Visual-Motor Integration and its subtests related to achievement test scores? *Optom. Vis. Sci.* **2003**, *80*, 758–763. [CrossRef]
2. Cameron, C.E.; Brock, L.L.; Murrah, W.M.; Bell, L.H.; Worzalla, S.L.; Grissmer, D.; Morrison, F.J. Fine motor skills and executive function both contribute to kindergarten achievement. *Child Dev.* **2012**, *83*, 1229–1244. [CrossRef] [PubMed]
3. Bender, L. *A Visual Motor Gestalt Test and Its Clinical Use*; American Orthopsychiatric Association: New York, NY, USA, 1938.
4. Koppitz, E.M. *The Bender-Gestalt Test for Young Children*; Grune & Stratton: New York, NY, USA, 1963.
5. Koppitz, E.M. *The Bender-Gestalt Test for Young Children: Research and Application, 1963–1973, Vol. 2*; Grune & Stratton: New York, NY, USA, 1975.
6. Santucci, H.; Galifret-Granjon, N. *Le Test Visuo-Moteur de Bender*; Éditions du Centre de Psychologie Appliquée: Paris, France, 1988.
7. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.
8. Reynolds, C.R.; Livingston, R.B.; Willson, V. *Measurement and Assessment in Education*; Pearson/Allyn & Bacon: Boston, MA, USA, 2006.
9. Ardila, A. Cultural values underlying psychometric cognitive testing. *Neuropsychol. Rev.* **2005**, *15*, 185–195. [CrossRef] [PubMed]
10. Rosselli, M.; Ardila, A. The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain Cogn.* **2003**, *52*, 326–333. [CrossRef] [PubMed]
11. Viljoen, G.; Levett, A.; Tredoux, C.; Anderson, S. Using the Bender Gestalt in South Africa: Some normative data for Zulu-speaking children. *S. Afr. J. Psychol.* **1994**, *24*, 145–151. [CrossRef]
12. Chan, P.W. Relationship of visual motor development and academic performance of young children in Hong Kong assessed on the Bender-Gestalt Test. *Percept. Mot. Ski.* **2000**, *90*, 209–214. [CrossRef] [PubMed]
13. Mazzeschi, C.; Lis, A. The Bender-Gestalt Test: Koppitz's Developmental Scoring System administered to two samples of Italian preschool and primary school children. *Percept. Mot. Ski.* **1999**, *88*, 1235–1244. [CrossRef] [PubMed]
14. Ghassemzadeh, H. A pilot study of the Bender-Gestalt Test in a sample of Iranian normal children. *J. Clin. Psychol.* **1988**, *44*, 787–792.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.