Vol. 7, No. 1, 2026, pp. 809-821
DOI: https://doi.org/10.29210/08jces695700

Contents lists available at Journal ELORA

Journal of Counseling, Education and Society
ISSN: 2716-4896 (Print) , ISSN 2716-4888 (Electronic)

Journal homepage: https://jces.eloracenter.org/jces



Development and initial validation of D-GRAFIMAT: an android-based error-pattern screening tool for mathematical dysgraphia in fourth-grade students

Puji Lestari*), Ati Rosliyati, Satya Santika, Sandy Ihsan Amarulloh
Universitas Siliwangi, Tasikmalaya, West Java, Indonesia

Article Info

Article history:

Received Sept 14th, 2026
Revised Oct 29th, 2026
Accepted Nov 21th, 2026

Keywords:

Mathematical dysgraphia,
Screening tool,
Content validity,
Error analysis,
Inclusive education,
Android application

ABSTRACT

Writing numbers, mathematical symbols, and computational steps is foundational to elementary school mathematics, yet some students who understand a concept orally fail to express it in writing. This difficulty, which the present study regards as a mathematics-specific manifestation of impairment in written expression (mathematical dysgraphia), is often mistaken for conceptual weakness because classroom instruments rarely separate writing errors from conceptual errors. This study developed D-GRAFIMAT, an Android-based screening tool that combines a 15-item written test designed to capture error patterns in numbers, symbols, and procedures with a structured four-aspect teacher observation sheet, and conducted its initial validation. The study followed the Borg and Gall research and development model, adapted into seven stages. The needs analysis involved 3 teachers, content validation involved seven experts, and a limited field trial involved 63 fourth-grade students from three intact classes in two public elementary schools in Tasikmalaya Regency, Indonesia. Content validity was high (Aiken's V = 0.86–1.00; S-CVI/Ave = 0.94), with excellent inter-expert agreement (ICC = 0.91). The internal consistency of the written test was adequate (Cronbach's α = 0.82), and teachers rated the tool as practical (89.5% of the maximum score). Of the 63 students, 87.3% were classified as normal, 11.1% as at risk, and 1.6% as indicated, and the tool reported a profile across the four aspects rather than a single score. D-GRAFIMAT is promising as a first-stage screening tool for teachers; its diagnostic accuracy against professional assessment has not yet been established and is the focus of further research.

© 2026 The Authors.
This is an open access article under the CC BY-NC-SA license
(https://creativecommons.org/licenses/by-nc-sa/4.0)

Corresponding Author:

Puji Lestari,
Universitas Siliwangi
Email: pujilestari@unsil.ac.id

Introduction

Inclusive education requires schools to recognize and respond to the learning needs of every student. This commitment is set out globally in Sustainable Development Goal (SDG) 4 and nationally in Indonesian regulations that require schools to provide reasonable accommodations for students with disabilities (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi [Kemendikbudristek], 2023). In practice, identification remains a weak point. Research in Indonesia has examined elementary teachers' competence in supporting students with learning difficulties and has documented recurring problems in implementing inclusive education, including in identifying students' needs (Azizah et al., 2024; Ediyanto et al., 2023). Writing difficulties are among the easiest to overlook. Handwriting difficulties are estimated to affect 10% to 30% of

school-age children (Feder & Majnemer, 2007; Mekyska et al., 2017), developmental writing disorders about 7% to 15% (Döhla & Heim, 2016), and dysgraphia remains common across all grade levels in certain groups of students (Mayes et al., 2019). Even so, dysgraphia has been studied far less than dyslexia (Döhla & Heim, 2016; Kalenjuk et al., 2022; McCloskey & Rapp, 2017) and often goes unnoticed by educators (Kalenjuk et al., 2024).

Dysgraphia refers to a disturbance in the production of written language related to the mechanics of writing (Biotteau et al., 2019). The DSM-5 does not list dysgraphia as a separate diagnosis: problems of written expression are classified under specific learning disorder with impairment in written expression, whereas handwriting problems related to motor coordination are associated with developmental coordination disorder (American Psychiatric Association, 2013; Biotteau et al., 2019; Chung et al., 2020). Handwriting depends on several component skills, including fine motor control, visual-motor integration, and visual perception (Feder & Majnemer, 2007), and dysgraphia can affect the written product, the writing process, or both (Biotteau et al., 2019). Accordingly, in this study *mathematical dysgraphia* is used as an operational term rather than a diagnostic category. It denotes manifestations of written-expression and graphomotor impairment in mathematical notation, such as reversed or malformed digits, interchanged operation signs, and misaligned place-value columns. This distinguishes it from dyscalculia, which concerns difficulties in processing numbers and calculating and is heterogeneous in nature (Geary, 2004; Kaufmann et al., 2013). The two frequently co-occur: among 1,006 students aged 6 to 16 years, dysgraphia was especially prevalent among students with mathematics learning disabilities (72%; Mayes et al., 2018). Because the same wrong answer can stem from either a conceptual error or a writing error, teachers need a tool that can separate the two.

Two bodies of research are relevant to this problem. The first is digital handwriting assessment. Tablet-based systems that record the kinematics, pressure, and tilt of handwriting have identified dysgraphia with 96% sensitivity and specificity (Mekyska et al., 2017), with accuracy comparable to expert raters using the BHK test (Asselborn et al., 2018), and with accuracies of about 73% to 80% in other samples (Devillaine et al., 2021; Drotár & Dobeš, 2020). Related studies have shown that mild dysgraphia often goes undetected in schools (Gargot et al., 2020) and that graphomotor difficulties can be quantified in third- and fourth-grade children (Galaz et al., 2020; Mucha et al., 2020). However, these systems analyze letters, words, or drawings rather than mathematical notation; they depend on digitizing tablets and specialist judgments to train their models; and they were developed outside Indonesia. The second body is Indonesian classroom research, which describes how teachers handle writing difficulties through observation and case-based mentoring (Sihombing et al., 2022) and discusses the potential of assistive technology for inclusive education (Suwahyo et al., 2022), but does not yet provide a validated screening instrument for mathematical writing. Guidance on handwriting assessment recommends combining standardized tools with classroom observation and teacher consultation (Feder & Majnemer, 2007), and the diversity of handwriting difficulties calls for multiple forms of classroom assessment (Prunty & Barnett, 2017). To our knowledge, no tool yet combines error-pattern analysis of mathematical writing with structured, multi-aspect teacher observation to produce an aspect-level profile that classroom teachers can use directly.

This study addresses that gap. Its contribution lies not in moving a paper test onto a phone but in its detection mechanism. D-GRAFIMAT pairs a written test whose items are designed to elicit specific writing-error patterns in numbers, symbols, and procedures with structured observation of four aspects derived from the motor, spatial, and cognitive components of handwriting (Biotteau et al., 2019; Feder & Majnemer, 2007; McCloskey & Rapp, 2017), and then reports a profile across these aspects. The tool is positioned as a screening instrument that flags students for further assessment; it does not diagnose. The study was conducted in two public elementary schools in Tasikmalaya Regency, West Java, where teachers reported students who could answer orally but could not write their solutions correctly. Grade IV was chosen because the curriculum at this level begins to introduce integers and simple fractions, which require multistep written procedures and careful notation.

Based on the foregoing, this study aimed to develop D-GRAFIMAT and conduct its initial validation. The research questions were: (1) What is the content validity of D-GRAFIMAT according to expert judgment? (2) What is its initial reliability, in terms of inter-expert agreement and internal consistency? (3) How practical is it according to teachers? (4) How are fourth-grade students distributed across the screening categories in a limited field trial, and what are the aspect-level profiles of the students who were flagged?

Method

Research design

This study used a research and development (R&D) design based on the Borg and Gall (1983) model. The model was chosen because it links product development to successive rounds of expert review and field testing, which suits an instrument that must be technically sound and usable by classroom teachers. Borg and Gall describe ten

steps, from information collection to dissemination and implementation. Because the aim of this study was the development and initial validation of a screening tool, not an evaluation of effectiveness or large-scale implementation, the study covered the steps up to the preliminary field test and the subsequent product revision, organized into seven stages (Table 1 and Figure 1). The main field test, operational field test, and dissemination (steps 6 to 10) require evidence that is not yet available, particularly diagnostic accuracy against a reference standard, and are planned for the next phase of research.

Table 1. Correspondence between the stages of this study and the Borg and Gall (1983) steps

Stage in this study	Borg and Gall (1983) step	Main output
Needs analysis	Research and information collecting	Problem description and candidate indicators
Instrument development	Planning; 3. Developing the preliminary form of the product	Draft written test, observation sheet, and test blueprint
Expert validation	Developing the preliminary form of the product (expert review)	Content validity indices; revised items
Application development	Developing the preliminary form of the product	Android prototype
Limited field trial	Preliminary field testing	Student scores, observation scores, teacher feedback
Data analysis	Preliminary field testing (evaluation)	Reliability, score distribution, practicality
Reflection and revision	Main product revision	Revised D-GRAFIMAT
Not conducted	Main field testing through dissemination and implementation	Planned for further research

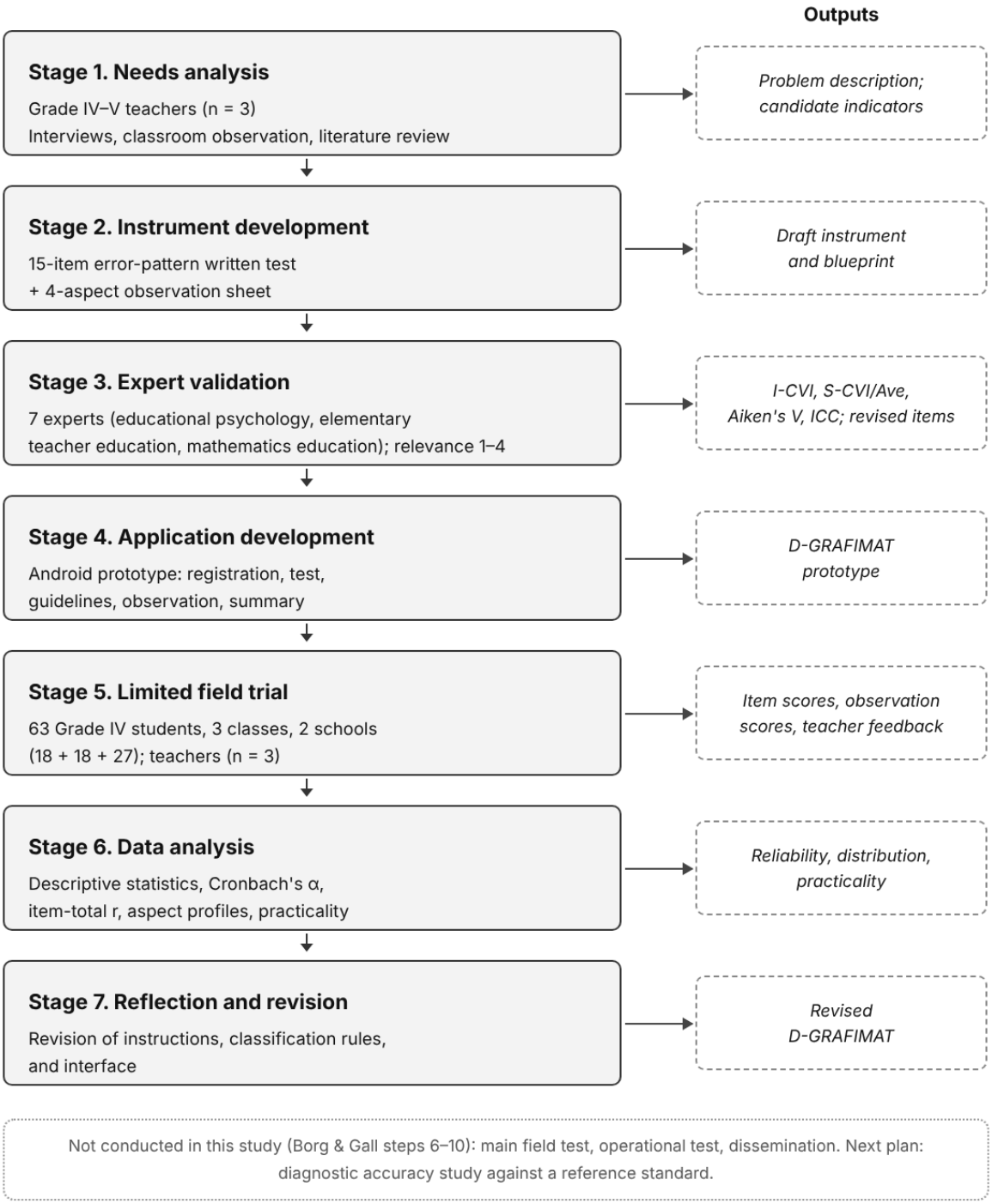


Figure 1. Research and development stages, participants, and outputs

Participants and research setting

The study was conducted in two public elementary schools in Tasikmalaya Regency, West Java, namely SDN 2 Kiarajangkung and SDN Sukahening. The participants at each stage are summarized in Table 2. Purposive sampling was used for teachers and experts. For the field trial, three intact fourth-grade classes were selected, and all students whose parents gave consent took part (N = 63). Because this was a convenience sample, the results are not intended to estimate prevalence.

Because this study did not test hypotheses, the sample size was not determined by power analysis. Lynn (1986) recommends three to ten experts for content validation; seven experts fall within this range and allow item-level indices to tolerate one divergent rating. For the field test, Borg and Gall (1983) recommend that the preliminary field test be conducted in a small number of schools (one to three) with a small number of subjects

(6 to 12). The 63 students from three classes exceed this minimum and are adequate for descriptive item analysis and an initial estimate of internal consistency, although not for precise estimation of classification accuracy.

Table 2. Participants at each research stage

Stage	Participants	n	Sampling technique	Inclusion criteria
Needs analysis	Grade IV and V teachers from both schools	3	Purposive	At least two years of teaching experience; experience handling students with writing difficulties
Expert validation	Lecturers in educational psychology (2), elementary teacher education (3), and mathematics education (2)	7	Purposive	Master's or doctoral degree in the field; at least five years of relevant experience
Limited field trial	Grade IV students: SDN 2 Kiarajangkung (18), SDN Sukahening IV A (18), SDN Sukahening IV B (27)	63	Intact classes (convenience)	All students in the selected classes whose parents gave consent [state any excluded students, if any]
Practicality evaluation	Classroom teachers who administered D-GRAFIMAT	3	Purposive	Teachers of the participating classes

Instrument: the two components of D-GRAFIMAT

D-GRAFIMAT consists of two components, both completed by a teacher or trained assessor and then entered into the application.

Component 1: error-pattern written test. The written test contains 15 items on whole numbers, integers, and simple fractions drawn from the Grade IV mathematics curriculum. The items are designed so that students who understand an operation may still display writing errors, for example by involving digits that are often reversed (such as 6 and 9), operation signs that are often interchanged (such as + and ×), and column procedures that require alignment of place-value columns. Table 3 presents the test blueprint. Each item is scored 0 (incorrect or illegible), 1 (partially correct, or conceptually correct but with errors in writing digits, symbols, or procedures), or 2 (correct and written correctly), so total scores range from 0 to 30.

Table 3. Blueprint of the error-pattern written test

Indicator	Targeted writing-error pattern	Item numbers	Number of items
Writing numbers (whole numbers, integers)	Reversed or malformed digits; e.g., 6/9, 2/5	1, 2, 3, 4	4
Writing operation and relation symbols	Interchanged +,×, −, =; omitted signs	5, 7, 10, 12	4
Writing multistep procedures	Misaligned place-value columns; omitted or reordered steps	6, 8, 9, 11	4
Writing simple fractions	Interchanged numerator/denominator; fraction-bar errors	13, 14, 15	3
Total		**1–15**	**15**

Component 2: structured observation sheet. While students work on the test, the assessor completes an observation sheet of 16 statements grouped into four aspects: (a) fine motor skills and handwriting, (b) spatial orientation and visual perception, (c) cognitive processes in writing, and (d) self-perception and subjective experience. Example statements are "The student has difficulty keeping the shape of digits consistent" for the fine motor aspect; "The student writes digits or operations out of alignment with the intended column" for the spatial orientation aspect; "The student skips steps when writing out a computational procedure" for the cognitive process aspect; and "The student shows hesitation or reluctance to write mathematical answers" for the self-perception aspect. Each statement is rated on a four-point frequency scale (1 = never, 2 = sometimes, 3 = often, 4 = always); higher scores indicate more frequent difficulty. The first two aspects reflect the motor and visuospatial components underlying handwriting (Biotteau et al., 2019; Feder & Majnemer, 2007), the third reflects the cognitive mechanisms of producing writing (McCloskey & Rapp, 2017), and the fourth reflects the reported impact of handwriting difficulties on self-esteem (Feder & Majnemer, 2007). Combining a test with structured classroom observation follows the recommendation that handwriting be evaluated with standardized tools accompanied by observation and teacher consultation (Feder & Majnemer, 2007).

Classification rules. The application computes the total test score and the mean score of each observation aspect, then assigns one of three screening categories: normal, at risk, or indicated. A student is classified as indicated if they obtain a test score of ≤18 out of 30 and a mean score of ≥3.00 on at least two observation aspects. A student is classified as at risk if they obtain a test score of 19–23 or a mean score of ≥2.50 on at least

two observation aspects but do not meet the indicated criteria. A student is classified as normal if they obtain a test score of ≥24 and do not meet the risk criteria on the observation. The cut-off scores were set by consensus of the seven experts after reviewing the blueprint, item characteristics, and sample student profiles. Categories are computed automatically from the scores entered by the assessor; the application does not analyze images of students' handwriting. The categories indicate the need for further assessment and do not constitute a diagnosis.

Procedure

Needs analysis. Semi-structured interviews were conducted with three Grade IV teachers from the two schools using a protocol of eight main questions about (a) the writing errors they observed in mathematics, (b) how they distinguished these errors from conceptual errors, and (c) the assessment tools available to them. Each interview lasted about 25–35 minutes and was complemented by classroom observation. Interview notes were analyzed thematically by two researchers; coding disagreements were resolved through discussion until consensus was reached. A literature review on dysgraphia, mathematics learning difficulties, and digital assessment informed the selection of indicators.

Expert validation. Each expert received the test items, observation sheet, blueprint, and validation sheet. The validation thus covered the relevance of the 15 test items and the appropriateness of the statements on the observation sheet. The experts rated the relevance of each test item to its indicator on a four-point scale (1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, 4 = highly relevant). The experts also rated five aspects of the instrument (alignment of content with indicators, item construction, suitability of language for elementary students, readability of instructions, and visual layout) on the same scale and wrote comments. Items were revised based on these indices and comments before being integrated into the application.

Application development. The validated instrument was implemented as an Android application using Android Studio with the Kotlin programming language. For the simulation in this manuscript, the application is assumed to be usable offline after installation; data storage and summaries are handled on the device, while data backup or export takes place when the device reconnects to the internet. The application workflow is sequential: student registration, written test, scoring guidelines, observation sheet, and results summary (Figures 2 to 6). The interface was designed for teachers with limited experience using digital applications.

Limited field trial. The tool was trialed in September 2026 by three classroom teachers who had been briefed on how to use the application. Each teacher administered the tool individually to students from their own class in a single session of about 35–45 minutes. Students wrote their answers on paper, and the assessor then entered the score for each item and the observation scores into the application. After the trial, the three teachers completed a 19-item practicality questionnaire on ease of use, clarity of guidelines, time required, and usefulness of the results, using a four-point scale (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree).

Reflection and revision. Expert comments, teacher feedback, and trial results were used to revise item instructions, classification rules, and the interface.

Data analysis

Content validity was analyzed at the item level. The item-level content validity index (I-CVI) was computed as the proportion of experts giving an item a rating of 3 or 4, and the scale-level index as the mean $I-CVI$ ($S-CVI/Ave$; Polit & Beck, 2006). Following Lynn (1986) and Polit et al. (2007), I-CVI ≥ 0.78 and S-CVI/Ave ≥ 0.90 were considered to indicate excellent content validity; with seven experts, an item rated relevant by six experts (I-CVI = 0.86) meets this criterion. Aiken's V was also computed for each item using the formula $V = \Sigma(r - 1) / [n(c - 1)]$, where r is an expert's rating, n = 7 is the number of experts, and c = 4 is the number of rating categories. These values were compared with the critical value from Aiken (1985) for seven raters and four categories ($V \geq 0.76$, $p < .05$).

Inter-expert agreement was estimated with the intraclass correlation coefficient (two-way random effects, absolute agreement, mean of k raters; Shrout & Fleiss, 1979) with a 95% confidence interval and interpreted following Koo and Li (2016): below 0.50 poor, 0.50 to 0.75 moderate, 0.75 to 0.90 good, and above 0.90 excellent. The internal consistency of the written test and of each observation aspect was estimated with $Cronbach's\ \alpha$, and corrected item-total correlations were computed to identify weak items. Because α depends on the number of items and score variance and does not demonstrate unidimensionality (Taber, 2018), its values were interpreted together with item statistics rather than against a single threshold.

Practicality scores were summarized as means and as percentages of the maximum score. For descriptive interpretation, percentages of 81–100% were categorized as very practical, 61–80% as practical, 41–60% as fairly practical, and ≤40% as less practical. Field trial results were summarized as descriptive statistics of the total test score and the number of students in each screening category. For students classified as at risk or indicated, aspect-level profiles were examined to determine which aspects contributed most to their classification. Expert comments and teacher notes were summarized descriptively. Sensitivity, specificity, and predictive values were not estimated because they require a reference-standard comparator, such as independent assessment by a

psychologist (Trevethan, 2017), which was beyond the scope of this initial study. Ethical considerations. Before data collection, the researchers obtained written permission from the principals of both schools. Parents provided written consent, students provided verbal assent, and all data were anonymized before analysis.

Results and Discussion

Needs analysis

The interviews with the three teachers revealed three main themes. First, all three teachers (3 of 3) had encountered students who could explain an answer orally but had difficulty writing it in the form of numbers, symbols, or mathematical procedures. Second, two teachers (2 of 3) stated that writing errors were often treated as conceptual errors because no guidelines were available to distinguish the two. Third, all three teachers (3 of 3) had not used any specific instrument to record patterns of mathematical writing errors systematically. The teachers also expressed a need for an instrument that is easy to use, has clear scoring guidelines, and presents results by aspect. These findings, together with the literature review, formed the basis for the two-component design described above.

Content validity and inter-expert agreement

Table 4 summarizes the expert ratings by aspect; item-level indices for the 15 test items are presented in Appendix A. Aiken's V for the 15 items ranged from 0.86 to 1.00, with $I-CVI = 0.86-1.00$ and an $S-CVI/Ave$ of 0.94, which exceeds the 0.90 criterion, and inter-expert agreement was excellent (ICC = 0.91; 95% CI 0.86–0.95). The lowest value was found for item 14 on writing fractions, which the experts judged to have relatively long instructions that students might interpret differently. In response, the instructions for the fraction items were simplified, ambiguous symbols were redrawn, and the layout was adjusted to improve readability.

Table 4. Content validity of the D-GRAFIMAT written test by aspect (seven experts)

Aspect	Mean score (SD)	Aiken's V (range across items)	I-CVI (range)	Interpretation
Alignment of content with indicators	3.66 (0.36)	0.86–1.00	0.86–1.00	Excellent
Item construction	3.58 (0.41)	0.86–1.00	0.86–1.00	Excellent
Suitability of language for elementary students	3.72 (0.32)	0.86–1.00	0.86–1.00	Excellent
Readability of instructions	3.64 (0.39)	0.86–1.00	0.86–1.00	Excellent
Visual layout (font, spacing)	3.51 (0.44)	0.86–1.00	0.86–1.00	Excellent
Overall (15 items)	3.62 (0.39)	0.86–1.00	S-CVI/Ave = 0.94	Excellent; ICC = 0.91 (95% CI 0.86–0.95)

These indices indicate that the experts judged the test items to be relevant to their indicators and appropriate for fourth-grade students. However, content validity is evidence only of the relevance and representativeness of the items; it does not show that the tool measures the intended construct or that its scores are reliable (Polit et al., 2007). Reliability is therefore examined separately below, while construct validity remains a task for further research.

Reliability

The internal consistency of the written test was adequate (α = 0.82), and corrected item-total correlations ranged from 0.24 to 0.68 (Table 5). In this simulation, no item had a corrected item-total correlation below 0.20; all items were retained, with wording refinements to the items that received expert comments. For the observation aspects, α ranged from 0.74 to 0.81. This level of consistency was achieved even though most students performed well (87.3% classified as normal), which restricts score variance and tends to lower α. Because α also depends on the number of items and does not demonstrate unidimensionality, these estimates should be regarded as preliminary and re-examined in a larger and more heterogeneous sample (Taber, 2018).

Table 5. Reliability indices of D-GRAFIMAT (N = 63 students; seven experts)

Component	Items	Cronbach's α	Corrected item-total r (range)
Error-pattern written test (score 0–30)	15	0.82	0.24–0.68
Observation: fine motor skills and handwriting	4	0.78	0.35–0.67
Observation: spatial orientation and visual perception	4	0.81	0.38–0.71
Observation: cognitive processes in writing	4	0.76	0.32–0.65

Component	Items	Cronbach's α	Corrected item-total r (range)
Observation: self-perception and subjective experience	4	0.74	0.30–0.62
Inter-expert agreement (ICC, two-way random effects, absolute agreement)	15	ICC = 0.91	95% CI 0.86–0.95

Practicality

Teachers rated D-GRAFIMAT as practical, with a mean practicality score of 89.5% of the maximum; the highest ratings were given to the clarity of the usage instructions and ease of navigation, and the lowest to the speed of summarizing results when data for several students were entered consecutively. All three teachers stated that the application workflow was easy to understand and that the aspect-level profile helped them determine appropriate follow-up. Two teachers suggested adding a feature for exporting results, and one teacher proposed adding an example of how to complete the observation sheet. Based on this feedback, the usage instructions and the example of completing the observation sheet were clarified.

Limited field trial results

Table 6 presents the distribution of screening results. Of the 63 students, 55 (87.3%) were classified as normal, 7 (11.1%) as at risk, and 1 (1.6%) as indicated. The total test score had a mean of 26.1 (SD = 3.2, range 17 to 30).

Table 6. Distribution of screening results in the limited field trial

School	Class	Number of students	Normal	At risk	Indicated
SDN 2 Kiarajangkung	IV	18	15	3	0
SDN Sukahening	IV A	18	16	1	1
SDN Sukahening	IV B	27	24	3	0
Total		**63**	**55 (87.3%)**	**7 (11.1%)**	**1 (1.6%)**

The aspect-level profiles of the eight flagged students varied. In the at-risk group, five of the seven students obtained relatively high observation scores on the spatial orientation and visual perception aspect, mainly in the form of misaligned place-value columns and inconsistent placement of digits. The other two at-risk students showed more prominent difficulties in the fine motor and handwriting aspect, such as inconsistent digit shapes and unstable pencil pressure. The one student in the indicated category showed high scores on the fine motor and spatial orientation aspects, accompanied by repeated errors in copying digits and setting out column procedures. These profiles illustrate the expected added value of the tool: two students in the same category may need different follow-up, for example fine motor practice for one and visual-spatial support, such as grid paper for aligning columns, for the other.

The proportion of flagged students (12.7% at risk or indicated) falls within the 10% to 30% range reported for handwriting difficulties in school-age children (Feder & Majnemer, 2007; Mekyska et al., 2017) and slightly above the estimated 5% to 10% for dysgraphia (Devillaine et al., 2021). This pattern is consistent with the role of a screening tool, which is expected to flag more students than will eventually receive a diagnosis. The proportion classified as indicated (one student, 1.6%) is small and, given the sample size, highly imprecise. These figures should not be read as prevalence estimates, because the sample came from three conveniently selected classes and the cut-off scores have not yet been calibrated against a reference standard.

The D-GRAFIMAT application

The validated instrument was integrated into an Android application that guides the assessor through a fixed sequence: student registration, written test, scoring guidelines, observation sheet, and results summary. The registration screen records the student's name, sex, class, school, and assessment date so that results can be linked to the correct student (Figure 2).

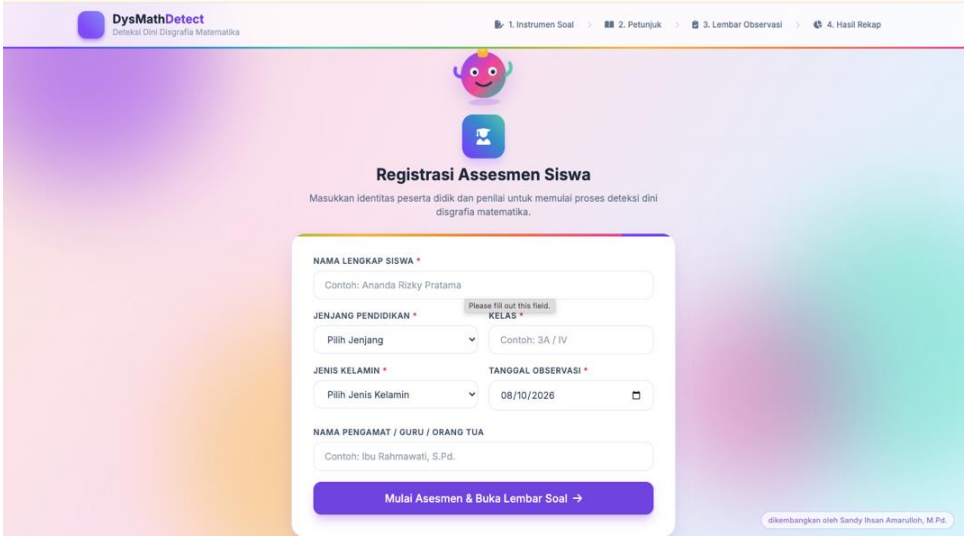


Figure 2. D-GRAFIMAT student registration interface

The assessor then administers the written test, whose items are presented with the same structure as the validated version (Figure 3), and refers to the scoring guidelines, which define the criteria for each score so that the assessor applies them consistently (Figure 4).

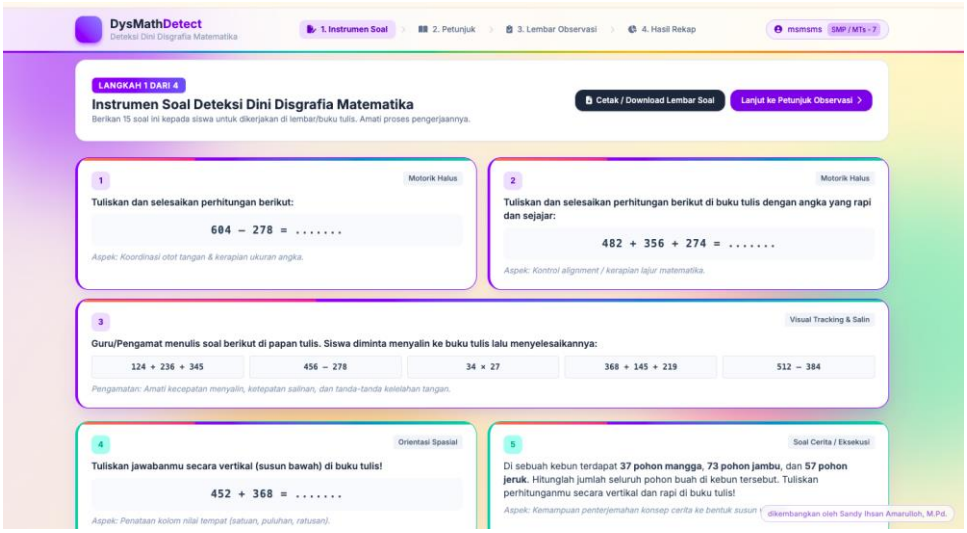


Figure 3. Error-pattern written test interface

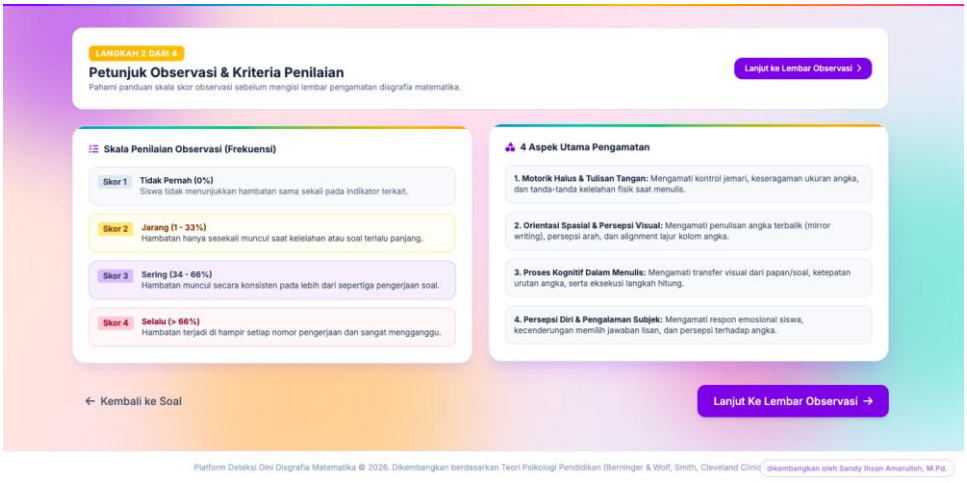


Figure 4. Scoring guidelines and criteria interface

The observation sheet presents the statements for the four aspects with a four-point frequency scale and a field for additional notes (Figure 5).

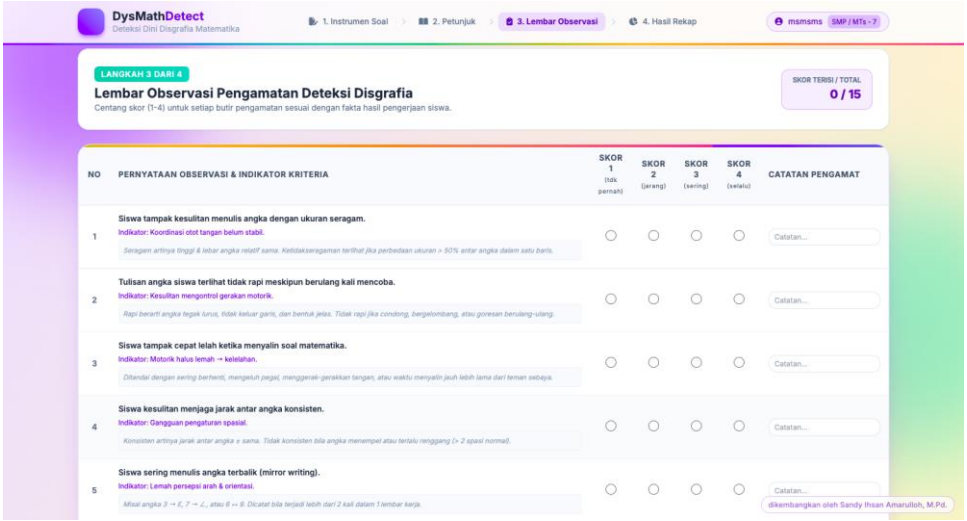


Figure 5. Observation sheet interface

The summary screen displays the total score, screening category, and a breakdown of the four aspects, followed by space for follow-up recommendations (Figure 6). Presenting aspect-level scores, rather than only the total score, allows the assessor to see where a student's difficulties lie. As noted earlier, the category is computed from the scores entered by the assessor; the application does not recognize handwriting automatically.

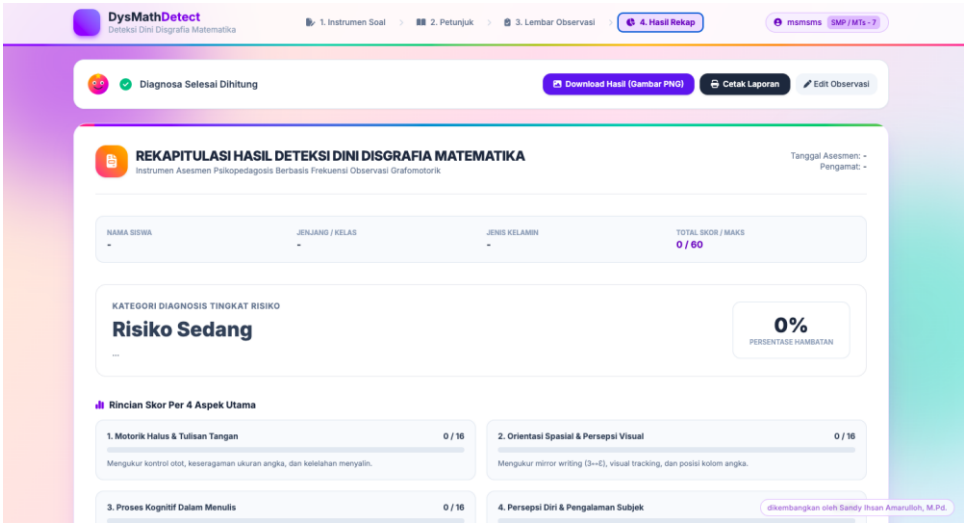


Figure 6. Results summary interface

Compared with tablet-based dysgraphia assessments that analyze the kinematics, pressure, and tilt of handwriting (Asselborn et al., 2018; Drotár & Dobeš, 2020; Mekyska et al., 2017), D-GRAFIMAT is technologically simple because it relies on assessor scoring rather than sensor data. Its strengths lie elsewhere. Those systems analyze letters, words, or drawings, whereas D-GRAFIMAT targets mathematical notation and is designed to separate writing errors from conceptual errors within the same task. The tool also runs on ordinary Android phones in classrooms that have no digitizing tablets. The consequence of this simplicity is that its accuracy depends on the consistency with which assessors apply the scoring guidelines, so inter-teacher agreement is an important target for further research.

The four observation aspects are consistent with the view that handwriting difficulties are heterogeneous and may involve motor execution, visuospatial organization, or higher-level writing processes (Biotteau et al., 2019; McCloskey & Rapp, 2017). Clustering of handwriting data has also identified distinct dysgraphia profiles rather than a single continuum of severity (Gargot et al., 2020). From the perspective of learning with multiple representations, students who can explain a procedure orally but cannot write it down have difficulty not with the concept but with translating that concept into a written symbolic representation (Ainsworth, 2006). Research

on student-generated representations likewise stresses that what students produce should be judged against the demands of the task (Ainsworth & Scheiter, 2021), and the D-GRAFIMAT written test is designed to reveal failures at this translation stage. Because writing and calculation difficulties frequently co-occur (Mayes et al., 2018) and dyscalculia is heterogeneous (Kaufmann et al., 2013), a low D-GRAFIMAT score should be followed up with assessment of both, rather than being attributed to writing alone.

For teachers, D-GRAFIMAT provides a structured way to follow up on the observation that a student "understands but cannot write," and the aspect-level profile can guide classroom accommodations while further assessment is pending. This is relevant for Indonesian elementary schools, where identifying students' needs remains a recurring problem in inclusive education (Azizah et al., 2024; Ediyanto et al., 2023) and regulations require schools to provide reasonable accommodations (Kemendikbudristek, 2023). The results should be used to refer students for further assessment and to plan support, not to label students.

This study has several limitations. First, the available evidence is limited to content validity, inter-expert agreement, internal consistency, and teacher-rated practicality; construct validity, test-retest reliability, and inter-teacher agreement have not yet been examined. Second, diagnostic accuracy has not been established because no reference-standard assessment was available; the screening categories therefore indicate risk, not diagnosis (Trevethan, 2017). Third, the field trial involved 63 students from three classes in two schools, which limits the generalizability and precision of the estimates. Fourth, classification depends on assessor scoring and has not been tested with teachers who were not involved in the development. The next phase will compare D-GRAFIMAT classifications with independent assessment by psychologists in a larger sample, estimate sensitivity, specificity, and predictive values with their confidence intervals, and examine inter-teacher agreement.

Conclusion

This study developed D-GRAFIMAT, an Android-based screening tool for mathematical dysgraphia that combines a 15-item error-pattern written test with a four-aspect teacher observation sheet, and conducted its initial validation with seven experts and 63 fourth-grade students. The written test showed high content validity (S-CVI/Ave = 0.94; Aiken's V = 0.86–1.00) with excellent inter-expert agreement (ICC = 0.91), its internal consistency was adequate ($\alpha = 0.82$), and teachers rated it as practical (89.5% of the maximum score). In the limited field trial, 87.3% of students were classified as normal, 11.1% as at risk, and 1.6% as indicated, and the flagged students showed varied aspect profiles. D-GRAFIMAT has the potential to help teachers distinguish writing errors from conceptual errors and to identify at-risk students who need further assessment; this potential still needs to be confirmed through a diagnostic accuracy study. Accordingly, D-GRAFIMAT is better described as a promising first-stage screening tool than as a diagnostic instrument. Before wider use, further research should evaluate its accuracy against professional assessment, its construct validity, and its inter-teacher reliability in a larger and more diverse sample of schools.

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

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Ainsworth, S. E., & Scheiter, K. (2021). Learning by drawing visual representations: Potential, purposes, and practical implications. *Current Directions in Psychological Science, 30*(1), 61–67. https://doi.org/10.1177/0963721420979582

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Asselborn, T., Gargot, T., Kidziński, Ł., Johal, W., Cohen, D., Jolly, C., & Dillenbourg, P. (2018). Automated human-level diagnosis of dysgraphia using a consumer tablet. *npj Digital Medicine, 1*, Article 42. https://doi.org/10.1038/s41746-018-0049-x

Azizah, N., Mumpuniarti, Rudiyati, S., & Evans, D. (2024). Elementary teachers' pedagogical competencies in supporting students with learning difficulties. *International Journal of Evaluation and Research in Education, 13*(2), 723–730. https://doi.org/10.11591/ijere.v13i2.26345

Biotteau, M., Danna, J., Baudou, É., Puyjarinet, F., Velay, J.-L., Albaret, J.-M., & Chaix, Y. (2019). Developmental coordination disorder and dysgraphia: Signs and symptoms, diagnosis, and rehabilitation. *Neuropsychiatric Disease and Treatment, 15*, 1873–1885. https://doi.org/10.2147/NDT.S120514

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). Longman.

Chung, P. J., Patel, D. R., & Nizami, I. (2020). Disorder of written expression and dysgraphia: Definition, diagnosis, and management. *Translational Pediatrics, 9*(Suppl. 1), S46–S54. https://doi.org/10.21037/tp.2019.11.01

Devillaine, L., Lambert, R., Boutet, J., Aloui, S., Brault, V., Jolly, C., & Labyt, E. (2021). Analysis of graphomotor tests with machine learning algorithms for an early and universal pre-diagnosis of dysgraphia. *Sensors, 21*(21), Article 7026. https://doi.org/10.3390/s21217026

Döhla, D., & Heim, S. (2016). Developmental dyslexia and dysgraphia: What can we learn from the one about the other? *Frontiers in Psychology, 6*, Article 2045. https://doi.org/10.3389/fpsyg.2015.02045

Drotár, P., & Dobeš, M. (2020). Dysgraphia detection through machine learning. *Scientific Reports, 10*, Article 21541. https://doi.org/10.1038/s41598-020-78611-9

Ediyanto, E., Ramadhani, R. S., Fitrasari, B. D., Kenila, E., Sunandar, A., Hastuti, W. D., & Suhendri, S. (2023). The problems in the implementation of inclusive education in primary schools. *Journal of ICSAR, 7*(1), 1–9. https://doi.org/10.17977/um005v7i12023p1

Feder, K. P., & Majnemer, A. (2007). Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology, 49*(4), 312–317. https://doi.org/10.1111/j.1469-8749.2007.00312.x

Galaz, Z., Mucha, J., Zvoncak, V., Mekyska, J., Smekal, Z., Safarova, K., Ondrackova, A., Urbanek, T., Havigerova, J. M., Bednarova, J., & Faundez-Zanuy, M. (2020). Advanced parametrization of graphomotor difficulties in school-aged children. *IEEE Access, 8*, 112883–112897. https://doi.org/10.1109/ACCESS.2020.3003214

Gargot, T., Asselborn, T., Pellerin, H., Zammouri, I., Anzalone, S. M., Casteran, L., Johal, W., Dillenbourg, P., Cohen, D., & Jolly, C. (2020). Acquisition of handwriting in children with and without dysgraphia: A computational approach. *PLOS ONE, 15*(9), Article e0237575. https://doi.org/10.1371/journal.pone.0237575

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*(1), 4–15. https://doi.org/10.1177/00222194040370010201

Kalenjuk, E., Laletas, S., Subban, P., & Wilson, S. (2022). A scoping review to map research on children with dysgraphia, their carers, and educators. *Australian Journal of Learning Difficulties, 27*(1), 19–63. https://doi.org/10.1080/19404158.2021.1999997

Kalenjuk, E., Subban, P., Laletas, S., & Wilson, S. (2024). 'That's not something that's necessarily on the radar': Educators' perspectives on dysgraphia. *Australasian Journal of Special and Inclusive Education, 48*(1), 1–16. https://doi.org/10.1017/jsi.2023.17

Kaufmann, L., Mazzocco, M. M., Dowker, A., von Aster, M., Göbel, S. M., Grabner, R. H., Henik, A., Jordan, N. C., Karmiloff-Smith, A. D., Kucian, K., Rubinsten, O., Szucs, D., Shalev, R., & Nuerk, H.-C. (2013). Dyscalculia from a developmental and differential perspective. *Frontiers in Psychology, 4*, Article 516. https://doi.org/10.3389/fpsyg.2013.00516

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2023). *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 48 Tahun 2023 tentang akomodasi yang layak untuk peserta didik penyandang disabilitas pada satuan pendidikan anak usia dini formal, pendidikan dasar, pendidikan menengah, dan pendidikan tinggi* [Regulation of the Minister of Education, Culture, Research, and Technology No. 48 of 2023 on reasonable accommodations for students with disabilities in formal early childhood, primary, secondary, and higher education]. https://peraturan.bpk.go.id/Details/285711

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

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Mayes, S. D., Breaux, R. P., Calhoun, S. L., & Frye, S. S. (2019). High prevalence of dysgraphia in elementary through high school students with ADHD and autism. *Journal of Attention Disorders, 23*(8), 787–796. https://doi.org/10.1177/1087054717720721

Mayes, S. D., Frye, S. S., Breaux, R. P., & Calhoun, S. L. (2018). Diagnostic, demographic, and neurocognitive correlates of dysgraphia in students with ADHD, autism, learning disabilities, and neurotypical development. *Journal of Developmental and Physical Disabilities, 30*(4), 489–507. https://doi.org/10.1007/s10882-018-9598-9

McCloskey, M., & Rapp, B. (2017). Developmental dysgraphia: An overview and framework for research. *Cognitive Neuropsychology, 34*(3–4), 65–82. https://doi.org/10.1080/02643294.2017.1369016

Mekyska, J., Faundez-Zanuy, M., Mzourek, Z., Galaz, Z., Smekal, Z., & Rosenblum, S. (2017). Identification and rating of developmental dysgraphia by handwriting analysis. *IEEE Transactions on Human-Machine Systems, 47*(2), 235–248. https://doi.org/10.1109/THMS.2016.2586605

Mucha, J., Mekyska, J., Galaz, Z., Faundez-Zanuy, M., Zvoncak, V., Safarova, K., Urbanek, T., Havigerova, J. M., Bednarova, J., & Smekal, Z. (2020). Analysis of various fractional order derivatives approaches in assessment of graphomotor difficulties. *IEEE Access, 8*, 218234–218244. https://doi.org/10.1109/ACCESS.2020.3042591

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467. https://doi.org/10.1002/nur.20199

Prunty, M., & Barnett, A. L. (2017). Understanding handwriting difficulties: A comparison of children with and without motor impairment. *Cognitive Neuropsychology, 34*(3–4), 205–218. https://doi.org/10.1080/02643294.2017.1376630

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sihombing, V. I. C., Rahman, R., & Damaianti, V. S. (2022). Upaya guru dalam mengatasi disgrafia (kesulitan menulis) pada siswa kelas rendah [Teachers' efforts to address dysgraphia (writing difficulties) in lower-grade students]. *Primary: Jurnal Pendidikan Guru Sekolah Dasar, 11*(3), 790–795. https://doi.org/10.33578/jpfkip.v11i3.8342

Suwahyo, B. W., Setyosari, P., & Praherdhiono, H. (2022). Pemanfaatan teknologi asistif dalam pendidikan inklusif [The use of assistive technology in inclusive education]. *Edcomtech: Jurnal Kajian Teknologi Pendidikan, 7*(1), 51–63. https://doi.org/10.17977/um039v7i12022p051

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*, Article 307. https://doi.org/10.3389/fpubh.2017.00307