

Feasibility and Preliminary Clinical Validation of the Explainable Digital Biomarker for Dyslexia (EDBD): A Prospective Pilot Study Protocol

Carrie VanZant

carrievanzant@icloud.com

Neuro Navigation Inc

Study protocol

Keywords: dyslexia, digital biomarker, explainable artificial intelligence, diagnostic accuracy, pilot study, school-based screening, STARD-AI, implementation science

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10768379/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Dyslexia affects an estimated 5–10% of school-age children, yet identification frequently occurs years after early warning signs emerge, delaying access to evidence-based intervention during the developmental window when remediation is most effective. Phases 1 through 4 of this translational research program synthesized four decades of dyslexia biomarker evidence, prioritized four evidence-anchored predictive markers, and specified the Explainable Digital Biomarker for Dyslexia (EDBD), a two-layer, clinician-facing screening and explanation architecture designed to flag children at elevated probability of dyslexia while making its reasoning auditable to non-specialist adults. No component of the EDBD has yet been tested with children.

Objective

This protocol describes a prospective, school-based pilot study designed to (1) evaluate the feasibility of administering the EDBD within real elementary school settings, (2) generate preliminary, hypothesis-generating estimates of its classification accuracy against a comprehensive psychoeducational reference standard, and (3) evaluate whether its explanation layer measurably supports, rather than undermines, clinician and educator judgment.

Methods

Following Standards for Reporting of Diagnostic Accuracy Studies (STARD) and STARD-AI guidance, we propose to enroll approximately 200 children in grades 1 through 3 from partnering elementary schools for universal EDBD screening, with a stratified sub-sample of approximately 60 children undergoing a blinded, comprehensive reference-standard evaluation. Feasibility will be assessed against pre-specified traffic-light progression criteria covering recruitment, completion, administration time, and technical reliability. Diagnostic accuracy indices (sensitivity, specificity, area under the curve, calibration) will be reported descriptively with confidence intervals appropriate to pilot-scale samples, without formal hypothesis testing. A nested user study will evaluate teacher and clinician trust, interpretability, and decision behavior using validated usability and trust instruments and Cohen's kappa. All procedures require Institutional Review Board approval, parental permission, and developmentally appropriate child assent.

Results

This is a study protocol; no data have been collected, and no outcomes are reported.

Conclusions

If feasibility and progression criteria are met, this pilot will directly inform the design of a fully powered, multi-site diagnostic accuracy trial and will provide the first empirical evidence base for responsibly deploying an explainable AI screening tool for early identification of neurodivergent learners.

Background

Dyslexia is the most common neurodevelopmental learning disability, affecting an estimated 5% to 10% of school-age children and accounting for a substantial share of the roughly 15% of the student population identified as neurodivergent or otherwise disabled under federal education law. Despite decades of behavioral, cognitive, and neurobiological research establishing dyslexia's early markers, the modal age of formal identification in the United States remains third grade or later, well past the kindergarten-to-first-grade window in which phonological, orthographic, and rapid-naming interventions appear most effective. This gap between what the science of early identification makes possible and what routinely happens in classrooms is the animating problem of this five-phase translational research program.

This protocol is the fifth and final phase of a program that began with a systematic review of four decades of dyslexia biomarker evidence spanning behavioral, eye-movement, neuroimaging, and genetic domains (Phase 1). That evidence was organized into a structured, queryable evidence atlas cataloguing the psychometric quality, cost, scalability, and evidence strength of every candidate marker identified in the literature (Phase 2). The atlas was then used to prioritize a small set of markers (phonemic awareness, rapid automatized naming [RAN], orthographic-choice accuracy, and serial-order verbal working memory) on the basis of predictive validity, feasibility of collection in a general-education classroom, and freedom from requiring specialized equipment or clinical personnel (Phase 3). Phase 4 translated this prioritized marker set into a full technical specification for the Explainable Digital Biomarker for Dyslexia (EDBD): a two-layer architecture in which an interpretable, additive statistical model (Layer 1) produces a probability estimate and confidence interval from the four markers, and a template-based reasoning and visualization layer (Layer 2) renders that probability, its contributing factors, and an explicit uncertainty band into a waterfall-style display intended for teachers, school psychologists, and pediatric clinicians who are not machine-learning specialists.

Each of the first four phases was a synthesis of existing evidence and a specification exercise; none involved collecting new data from children. This is a deliberate and, we argue, necessary sequencing: a screening tool intended for use with a vulnerable pediatric population should not be piloted in classrooms before its architecture, evidence base, and explanation design have been fully specified and internally audited. Having completed that specification, the program now reaches the point at which further progress requires empirical contact with the setting and population the EDBD is meant to serve. The purpose of this protocol is to describe, prospectively, how that empirical contact will be made

responsibly, incrementally, and in a manner that generates genuine evidence rather than premature claims of validity.

Two distinct classes of uncertainty motivate this pilot. The first is operational: it is not yet known whether the EDBD's four-marker battery can be administered within a typical elementary school's scheduling, staffing, and technology constraints, whether children in the target age range can complete the tasks as designed, or whether the resulting data are of sufficient quality to feed the Layer 1 model. The second is inferential: because the EDBD's Layer 1 weights were derived from prioritization of the published literature rather than fit to a dataset the EDBD itself collected, it is not yet known how its probability estimates compare to a comprehensive, in-person psychoeducational evaluation, the current clinical reference standard for a dyslexia determination, nor whether its Layer 2 explanations help or hinder the humans who must ultimately act on its output. A study that only tested accuracy would risk deploying an unusable tool; a study that only tested usability would risk validating a tool with no diagnostic value. This protocol addresses both.

Because the EDBD is fundamentally a diagnostic-accuracy technology, this protocol is structured according to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015 framework and its artificial-intelligence-specific extension, STARD-AI, which together specify the design, blinding, and reporting elements needed for a diagnostic accuracy study, including one involving an AI-based index test, to be interpretable and trustworthy [1,2,3]. We adopt STARD-AI's recommendation that AI-centered diagnostic studies report dataset provenance, algorithmic bias considerations, and human-AI interaction outcomes alongside conventional accuracy metrics. At the same time, because this is explicitly a pilot rather than a definitive validation trial, its statistical design follows contemporary guidance on pilot and feasibility studies, which cautions against treating small-sample accuracy estimates as confirmatory and instead recommends pre-specified, descriptive progression criteria to determine readiness for a fully powered trial [4,5,6].

This pilot has three co-equal, pre-specified objectives, none of which is intended to establish clinical-grade diagnostic accuracy on its own:

1. Feasibility. To determine whether universal EDBD screening can be delivered within general-education classrooms with acceptable rates of recruitment, task completion, administration time, and technical reliability, using pre-specified traffic-light progression criteria.
2. Preliminary diagnostic accuracy. To generate descriptive, hypothesis-generating estimates of the EDBD's sensitivity, specificity, discrimination (area under the receiver operating characteristic curve), and calibration against a comprehensive, blinded psychoeducational reference-standard evaluation, reported with confidence intervals appropriate to the sample size and explicitly not powered for a definitive accuracy claim.
3. Explainability and human factors. To evaluate whether the EDBD's Layer 2 explanation interface measurably affects the trust, interpretability judgments, and classification agreement (via Cohen's

kappa) of teachers and clinicians who review its output, and whether the explanation layer mitigates or exacerbates automation bias relative to a probability score presented without explanation.

This protocol treats rigor and humility as inseparable: the EDBD was designed to give families and teachers an earlier, more legible signal than they currently have, and whether it delivers on that promise is an empirical question this pilot is designed to begin answering, not one this program has any interest in assuming the answer to in advance.

Method

This is a prospective, single-arm, two-stage pilot study combining (a) a feasibility evaluation of universal EDBD screening across full classrooms and (b) a nested, blinded diagnostic-accuracy sub-study in which a stratified subset of screened children receives a comprehensive reference-standard evaluation. This nested case-control-enriched design, oversampling screen-positive children while randomly sampling a comparison group of screen-negative children for full evaluation, is a standard and resource-efficient approach in screening validation research, allowing preliminary accuracy estimation without requiring every participant to undergo the full multi-hour reference-standard battery [7,8]. The design is cross-sectional with respect to the index test and reference standard, which will be administered within a bounded time window to minimize maturation effects, consistent with STARD Item 22 guidance on time intervals between index test and reference standard [1].

The study will be conducted in partnering elementary schools recruited through the research team's existing educator and family network. Schools will be selected to reflect variation in socioeconomic composition and prior exposure to universal literacy screening, since implementation research on school-based screening consistently finds that resource constraints, screener training quality, and prior screening experience shape both completion rates and data quality [9,10]. At least two participating sites are planned so that feasibility findings are not artifacts of a single school's staffing or scheduling idiosyncrasies.

Eligible participants are children enrolled in grades 1 through 3 (approximate ages 6 to 9) at a participating school whose parent or legal guardian provides written informed consent and who themselves provide developmentally appropriate assent. This age range was selected because it spans the period in which the four EDBD markers (phonemic awareness, RAN, orthographic-choice accuracy, and serial-order verbal working memory) have the strongest published predictive validity for later reading difficulty while remaining early enough that identification could still meaningfully change the timing of intervention [11,12]. Children will be excluded from the diagnostic-accuracy sub-study (though not from the feasibility-only screening) if they have a documented uncorrected vision or hearing impairment that would confound task performance, an intellectual disability that would invalidate the reference-standard comparison, or insufficient English proficiency to complete English-language study measures, consistent with prior universal-screening validation work that has flagged English-learner status as a factor requiring separate psychometric consideration rather than exclusion from screening itself [13]. Non-

English-dominant children will still be screened for feasibility purposes, and their data will be analyzed as a pre-specified subgroup rather than discarded, in keeping with equity-in-screening recommendations [13].

Consistent with methodological guidance that pilot and feasibility studies should not be powered for formal hypothesis testing of clinical or diagnostic-accuracy outcomes, this protocol does not specify a sample size derived from a power calculation for sensitivity or specificity [4,6,14]. Instead, the target enrollment of approximately 200 children for universal screening is justified on feasibility grounds: it is large enough to span multiple classrooms and at least two schools, allowing meaningful estimation of recruitment and completion rates with reasonably narrow confidence intervals, while remaining achievable within a single school year given the program's staffing. For the nested diagnostic-accuracy sub-study, we plan to invite approximately 60 children: all children who screen positive (anticipated to be roughly 15–20% of the sample, consistent with published dyslexia prevalence estimates) plus a random sample of screen-negative children in a target ratio of approximately 1:1 to 2:1 negative-to-positive, for the full reference-standard evaluation. This sub-sample size draws on the widely cited rule of thumb that approximately 12 participants per group provides a reasonable basis for estimating the parameters (means, variances, or proportions) needed to design a subsequent, fully powered trial, while explicitly not claiming adequate power to test a formal accuracy hypothesis [15]. Reported sensitivity, specificity, and predictive values from this sub-study will therefore be accompanied by wide confidence intervals and explicitly framed as preliminary.

Because a central purpose of this pilot is to determine whether a larger, fully powered validation trial is warranted, and if so, what design changes it requires, we pre-specify quantitative feasibility progression criteria using a three-zone (red/amber/green) framework, an approach with growing precedent in pilot trial methodology for making the decision to proceed transparent and falsifiable rather than post hoc [16,5,17]. These criteria, along with the primary feasibility outcomes they are attached to, are presented in Table 1.

The System Under Evaluation: The EDBD Framework

The Explainable Digital Biomarker for Dyslexia, fully specified in Phase 4 of this program, is a two-layer clinical decision-support architecture. Layer 1 is an interpretable, additive statistical model that combines four evidence-anchored markers (phonemic awareness, rapid automatized naming, orthographic-choice accuracy, and serial-order verbal working memory) into a single probability estimate accompanied by a conformal-prediction-derived confidence interval, chosen specifically because additive, generalized-linear-model-style architectures preserve a direct, auditable relationship between each marker and the final score, in contrast to opaque ensemble or deep-learning approaches. Layer 2 renders that probability, its per-marker contributions, and its uncertainty band as a waterfall visualization, together with a template-based natural-language explanation and one of three recommended action bands: monitor, consider further evaluation, or pursue comprehensive evaluation. The Phase 4 specification incorporated explicit automation-bias safeguards, including uncertainty display and a

design principle that the tool must never present a bare classification without its contributing evidence, motivated by empirical findings that erroneous decision-support advice can measurably shift clinical decisions even among experienced practitioners, and that explanation quality (not mere presence) determines whether decision-support explanations reduce or amplify overreliance [18,19,20]. This pilot is the first empirical test of whether that architecture performs as designed when it leaves the specification document and enters an actual classroom.

Reference Standard

Children enrolled in the diagnostic-accuracy sub-study will receive a comprehensive, individually administered psychoeducational evaluation conducted by a licensed school psychologist or equivalently credentialed evaluator, following standard practice for establishing a dyslexia determination: standardized measures of word-reading accuracy and fluency, phonological processing, and, where indicated, cognitive ability, integrated with parent- and teacher-report data on academic history. This evaluation constitutes the reference standard against which EDBD Layer 1 output will be compared. Consistent with STARD guidance on avoiding incorporation bias, the evaluator conducting the reference-standard assessment will be blinded to the child's EDBD screening result at the time of evaluation, and EDBD scores will not be released to families, teachers, or evaluators until the reference-standard evaluation is complete and independently scored [1,2].

Procedure

Following IRB approval at the affiliated academic partner and at each participating school district, study staff will present the study to school administrators and teachers, distribute study information and consent materials to all eligible families, and obtain written parental or guardian consent prior to any contact with a child. Children whose parents consent will then be asked for their own assent using an age-appropriate script and visual aid, administered separately from the parental consent process, consistent with guidance that assent should be an ongoing, developmentally sensitive conversation rather than a single signature-equivalent event [21,22,23]. Because this study involves the collection of pediatric behavioral data intended, in part, to train and refine an AI system, procedures will additionally follow the ACCEPT-AI framework's recommendations for the responsible inclusion of pediatric data in AI/ML research, which addresses age-appropriate consent and assent, data minimization, and downstream data protection specifically for this use case [24].

All consenting, assenting children will complete the four-marker EDBD screening battery, administered via tablet by trained study staff (not the children's own classroom teachers, to reduce demand characteristics) in a quiet space at their school, in a single session targeted at 20 to 30 minutes. Administration time, task completion, technical errors, and child affect during the session will be logged by the administrator using a structured observation form. Teachers will separately complete a brief post-screening feasibility questionnaire addressing scheduling burden and perceived classroom disruption.

Following screening, children selected for the nested diagnostic-accuracy sub-study (all screen-positive children and a randomly sampled group of screen-negative children) will be scheduled for the comprehensive reference-standard evaluation, to be completed within four weeks of the EDBD screening session to limit the influence of instructional exposure or maturation on the comparison, an interval chosen to balance STARD's preference for a short index-test-to-reference-standard interval against the practical scheduling constraints of school-based licensed evaluators [1].

Nested Explainability and Trust Sub-Study

A separate sample of approximately 15 to 20 educators and clinicians (school psychologists, reading specialists, classroom teachers, and pediatric clinicians recruited through the research team's professional network) will participate in a structured, within-subjects user study modeled on published clinical-XAI evaluation designs [25,26]. Each participant will review a series of de-identified, simulated case vignettes constructed from realistic but non-patient-derived marker profiles (never real participant data) under two conditions presented in counterbalanced order: EDBD output with the full Layer 2 explanation interface, and EDBD output showing only the bare probability score and action band without supporting explanation. For each vignette, participants will record their own clinical judgment (whether they would recommend further evaluation), their confidence in that judgment, and their agreement or disagreement with the EDBD's recommendation. Agreement between participants' judgments and the EDBD's output, and between participants themselves, will be summarized using Cohen's kappa, a standard metric for inter-rater and human-AI classification agreement in comparable clinical-AI evaluation studies [27,28]. Participants will then complete the System Usability Scale (SUS) and a validated clinical-AI trust instrument, and take part in a brief semi-structured interview, following the mixed-methods pattern established in prior clinician-facing explainable-AI evaluations [25,29]. This sub-study uses simulated vignettes rather than live participant cases specifically so that explainability and trust can be evaluated on their own terms, independent of and prior to any claim about the EDBD's real-world diagnostic performance.

Outcome Measures

Primary Feasibility Outcomes

Consistent with core guidance that pilot studies should specify feasibility, not efficacy, as the primary outcome domain [30,14], the primary outcomes of this pilot are: (1) recruitment rate, defined as the proportion of eligible children whose families provide consent; (2) completion rate, defined as the proportion of consented, assenting children who complete the full EDBD screening battery; (3) administration time, defined as the wall-clock duration of the screening session; and (4) technical reliability, defined as the proportion of sessions completed without a device, connectivity, or software failure requiring session restart. Each outcome is attached to a pre-specified progression criterion, detailed in Table 1.

Preliminary Diagnostic Accuracy Outcomes

Within the nested diagnostic-accuracy sub-study, we will descriptively estimate sensitivity, specificity, positive and negative predictive value, and the area under the receiver operating characteristic curve (AUC) of the EDBD Layer 1 probability score against the reference-standard dyslexia determination, together with calibration (the correspondence between the EDBD's stated probability and the observed proportion of children with a positive reference-standard result), following contemporary calibration-reporting guidance for clinical prediction models [31]. Because the sub-sample size was chosen for feasibility and parameter-estimation purposes rather than statistical power, all accuracy estimates will be reported with exact (Clopper-Pearson) or Wilson score confidence intervals appropriate to small samples, and will not be used to support any claim of established diagnostic accuracy. No p-values or significance tests will be applied to these estimates, in keeping with recommendations against using pilot-scale data for confirmatory efficacy claims [6,14].

Explainability and Human-Factors Outcomes

Secondary outcomes from the nested trust sub-study include: SUS score for the Layer 2 interface; scores on the clinical-AI trust instrument; Cohen's kappa between participant judgment and EDBD recommendation, compared across the explained and unexplained conditions; the proportion of vignettes in which participants deferred to an EDBD recommendation they had initially disagreed with (an operational proxy for automation bias, following the framing used in prior clinical decision-support literature); and qualitative themes from the semi-structured interviews regarding perceived clarity, actionability, and appropriateness of the explanation content [36,20].

Ethical Considerations

This protocol will be submitted for review to the Institutional Review Board (IRB) of the affiliated academic research partner prior to any recruitment activity, and to each participating school district's research review process where applicable. No study procedures will begin at a given site until both approvals are obtained.

Because all participants are minors, this study requires a dual-layer permission process: written parental or legal-guardian informed consent, followed by separate, developmentally appropriate child assent obtained using age-calibrated language and, for younger children, visual aids. Consistent with the 2024 Declaration of Helsinki's emphasis on treating consent as an iterative, ongoing process rather than a one-time event, study staff will re-confirm assent verbally at the start of each session and will honor a child's dissent, even after parental consent and prior assent, as grounds for immediate, non-punitive withdrawal from that session [32,33,23]. Because the study involves the collection of behavioral data intended to inform an AI system, procedures follow the ACCEPT-AI framework's specific guidance on age-appropriate assent, communication, and equity for pediatric AI/ML research [24].

The anticipated risk level of this study is minimal: participation involves standardized educational screening tasks similar to those already administered in many schools, and a comprehensive evaluation that mirrors standard-of-care psychoeducational assessment practice, conducted by licensed

evaluators. The principal foreseeable risks are the emotional discomfort some children may experience during evaluation tasks that reveal reading difficulty, and the possibility that a screening result could be misused or misunderstood if not communicated carefully. To mitigate the first risk, study staff will be trained to monitor for child distress and to pause or discontinue any session on request. To mitigate the second, no individual result will be shared with a family or teacher without accompanying interpretive guidance from a study clinician, and screening results explicitly will not be used, on their own, to make any special-education eligibility determination, a use for which the EDBD was not designed and for which this pilot provides no evidentiary basis.

All child data will be de-identified for analysis and stored in compliance with the Family Educational Rights and Privacy Act (FERPA) and, where applicable, the Children's Online Privacy Protection Act (COPPA), using coded identifiers rather than names in any dataset used for EDBD analysis or refinement. Consistent with published findings that families and children conceptualize consent for health and educational data as an ongoing, renegotiable relationship rather than a one-time transaction, the parental consent form will specify the intended uses of collected data, retention period, and the family's right to withdraw data from future analysis, and will avoid representing data-sharing protections in more absolute terms than can be guaranteed [34].

Recruitment will explicitly target schools and families across a range of socioeconomic and linguistic backgrounds rather than a single high-resource pilot site, in direct response to documented evidence that algorithmic health and education tools risk performing unevenly across populations when development and validation samples are drawn from narrow, high-resource contexts [35,13]. Recruitment materials will be offered in the dominant home languages of participating school communities, and the study budget includes provision for translation and interpretation support during the consent process, reflecting this program's commitment to serving the full national population of neurodivergent learners rather than a demographically narrow subset of it.

Statistical Analysis Plan

Feasibility outcomes will be summarized descriptively (proportions and means with 95% confidence intervals) and compared against the pre-specified progression criteria in Table 1. Diagnostic accuracy outcomes (sensitivity, specificity, predictive values, AUC) will be estimated descriptively within the nested sub-sample using standard formulas with small-sample-appropriate confidence intervals, and calibration will be assessed visually via a calibration plot comparing predicted probability deciles to observed reference-standard-positive proportions, following standard calibration-reporting practice [31]. No formal hypothesis tests of diagnostic accuracy will be conducted, consistent with pilot-study methodology guidance that treatment- or accuracy-effect estimation, not hypothesis testing, is the appropriate statistical posture for underpowered pilot samples [6]. Trust and usability outcomes will be summarized descriptively, with within-subjects comparison of the explained versus unexplained conditions using appropriate non-parametric tests (e.g., Wilcoxon signed-rank) given the small planned sample, interpreted as exploratory. Qualitative interview data will be analyzed using thematic analysis. All

planned analyses are exploratory and hypothesis-generating; the study is not designed or powered to test a pre-specified diagnostic-accuracy hypothesis, and any observed point estimates will be reported alongside this limitation rather than as confirmatory findings.

Progression Criteria to a Full-Scale Validation Trial

Table 1 presents the pre-specified feasibility progression criteria. A “green” outcome across all four feasibility domains will be interpreted as support for proceeding directly to a fully powered, multi-site diagnostic accuracy trial. Any “amber” outcome will trigger a defined protocol modification (e.g., revised scheduling procedures, additional technical support) before proceeding. Any “red” outcome will require the EDBD architecture or study procedures to be substantively revised and, if warranted, re-piloted before a larger trial is considered, consistent with the traffic-light framework’s intended function of making the decision to proceed, or not, an evidence-based and falsifiable one rather than a foregone conclusion [16,5].

Feasibility Domain	Outcome Measure	Green (Proceed)	Amber (Modify)	Red (Reconsider)
Recruitment	% of eligible families providing consent	≥ 60%	40–59%	< 40%
Completion	% of consented, assenting children completing the full screening battery	≥ 90%	75–89%	< 75%
Administration time	Mean wall-clock duration of the screening session	≤ 30 min	31–40 min	> 40 min
Technical reliability	% of sessions completed without a device, connectivity, or software failure requiring restart	≥ 95%	85–94%	< 85%

Table 1. Pre-Specified Feasibility Progression Criteria

Anticipated Significance

This protocol completes the arc of a five-phase program that began with a systematic review of the dyslexia biomarker literature and ends, deliberately, not with a claim of validation but with a plan to earn one. That sequencing reflects a considered position about how AI-based tools intended for use with children should be developed: evidence synthesis and architectural specification first, followed by staged, transparently reported empirical testing, rather than simultaneous claims of both novelty and proven accuracy. If this pilot’s feasibility and preliminary accuracy findings support progression, the resulting fully powered trial will be the first study in the program to generate genuinely new data about children rather than to synthesize, prioritize, or specify based on existing publications, a threshold this

program has treated as consequential enough to warrant the full ethical and methodological infrastructure described above.

Beyond its direct output, this pilot is designed to contribute to two literatures the research team has identified as underdeveloped: responsible AI implementation science in education, and explainable-AI evaluation methods for pediatric and school-based settings. Much of the existing clinical-XAI evaluation literature has been conducted in adult, hospital-based decision-support contexts [37,36,20]; comparatively little work has evaluated whether explanation interfaces function similarly when the end users are classroom teachers and school psychologists rather than physicians, and when the subjects of the prediction are children rather than adult patients. The nested trust and explainability sub-study described here is intended to begin filling that gap, independent of whatever the EDBD's ultimate diagnostic accuracy turns out to be.

Limitations

Several limitations are anticipated and are stated here prospectively rather than deferred to a future results paper. First, the diagnostic-accuracy sub-sample is explicitly not powered to produce a definitive estimate of sensitivity or specificity; any point estimates this pilot generates should be understood as hypothesis-generating inputs to the design of a larger trial, not as evidence of validated clinical accuracy, and we commit to reporting them with this framing regardless of how favorable or unfavorable they prove to be. Second, recruitment through the research team's existing school and family network, while intentionally diversified across at least two sites, may still under-represent families with the least access to information about research participation, a limitation common to convenience-based pilot recruitment that a subsequent multi-site trial would need to address through more structured, population-representative sampling. Third, the nested case-control-enriched sub-sampling design, while standard in screening validation research, means that the sub-study's estimated predictive values (though not sensitivity and specificity) will not directly generalize to the true population prevalence of dyslexia and must be interpreted accordingly. Fourth, because the explainability sub-study uses simulated rather than live case vignettes, its findings speak to how clinicians and educators respond to the EDBD's explanation design in the abstract, not to how they would behave when facing a real, high-stakes decision about an actual child; this tradeoff was accepted deliberately, to allow explainability evaluation to proceed without exposing real children's data to a system whose accuracy has not yet been established, but it remains a real limitation on ecological validity. Finally, this protocol describes procedures only for children in grades 1 through 3; feasibility and accuracy in younger (pre-kindergarten) or older (upper elementary and beyond) populations, where the four EDBD markers may behave differently, remain outside its scope and are noted here as a direction for future work rather than a gap in the present design.

Conclusion

This protocol describes a prospective, staged pilot study designed to generate the first empirical evidence (feasibility, preliminary diagnostic accuracy, and human-factors performance) for the

Explainable Digital Biomarker for Dyslexia specified in Phase 4 of this program. Consistent with STARD and STARD-AI reporting standards and contemporary pilot-study methodology, the study is explicitly structured to avoid overclaiming from underpowered data, using pre-specified progression criteria to determine whether, and how, a fully powered validation trial should follow. No results are reported here because none yet exist; this document exists to make the path from specification to evidence transparent, falsifiable, and answerable to the population it is ultimately meant to serve. Completing this fifth phase marks the transition of the EDBD from a synthesis of what the literature says is possible to a plan for finding out, with real children, real teachers, and real clinicians, whether it actually works.

Declarations

Funding: This study protocol has not received external funding. Operational support is provided by Neuro Navigation, Inc.

Human Ethics and Consent to Participate declarations: This is a study protocol; no data collection has yet occurred and no human participants have been enrolled. Ethics review and consent procedures are described within the protocol and will be completed prior to any recruitment activity, in accordance with the 1964 Declaration of Helsinki and its subsequent amendments. IRB approval and written participant consent (and child assent where applicable) will be obtained before the study commences.

Competing interests: The author declares no competing interests. Neuro Navigation, Inc. is the organization developing the EDBD instrument described in this protocol; the author is the organization's Executive Director.

Author contributions: C.V. is the sole author and is responsible for all aspects of this work, including conceptualization, study design, methodology, protocol development, formal analysis framework, writing (original draft preparation, review, and editing), and project administration.

Data availability: Not applicable. This is a study protocol; no data have been collected.

References

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527.
2. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. *BMJ Open*. 2016;6(11):e012799.
3. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283–3289.

4. Kunselman AR. A brief overview of pilot studies and their sample size justification. *Fertil Steril*. 2024;121(6):899–901.
5. Lewis M, Bromley K, Sutton CJ, McCray G, Myers HL, Lancaster GA. Determining sample size for progression criteria for pragmatic pilot RCTs: The hypothesis test strikes back! *Pilot Feasibility Stud*. 2021;7:40.
6. Sim J. Should treatment effects be estimated in pilot and feasibility studies? *Pilot Feasibility Stud*. 2019;5:107.
7. Graney SB, Martinez RS, Missall KN, Aricak OT. Universal screening of reading in late elementary school: R-CBM versus CBM maze. *Remedial Spec Educ*. 2010;31(5):368–377.
8. Klingbeil DA, Nelson PM, Van Norman ER, Birr C. Diagnostic accuracy of multivariate universal screening procedures for reading in upper elementary grades. *Remedial Spec Educ*. 2017;38(5):308–320.
9. Ozernov-Palchik O, Elizee Z, Gabrieli JDE. (Not so) universal literacy screening: A survey of educators reveals variability in implementation. *Ann Dyslexia*. 2025;75:489–523.
10. Tatel S, Darcy LJ, Solari EJ. Determinants to implementing a new early literacy screener: Barriers and facilitators. *Ann Dyslexia*. 2025;75:524–546.
11. Schöfl M, Steinmair G, Holzinger D, Weber C. Predicting word reading deficits using an app-based screening tool at school entry. *Front Pediatr*. 2022;10:863477.
12. Zugarramurdi C, Fernández L, Lallier M, Carreiras M, Valle-Lisboa JC. Lexiland: A tablet-based universal screener for reading difficulties in the school context. *J Educ Comput Res*. 2022;60(7):1688–1715.
13. Siebert JM, Bell P, Gutiérrez N, Zegers M, Falke E, Domingue BW, et al. Differences in reading screening accuracy by percentile cutoff and English proficiency: Feature selection and group-wise prediction evaluation. *Read Res Q*. 2025;61(1):e70074.
14. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. *Med Care*. 2022;60(1):95–103.
15. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat*. 2005;4(4):287–291.
16. Herbert E, Julious SA, Goodacre S. Progression criteria in trials with an internal pilot: An audit of publicly funded randomised controlled trials. *Trials*. 2019;20:493.
17. Wilson DT, Hudson E, Brown S. Three-outcome designs for external pilot trials with progression criteria. *BMC Med Res Methodol*. 2024;24:226.
18. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19(1):121–127.
19. Goddard K, Roudsari A, Wyatt JC. Automation bias: Empirical results assessing influencing factors. *Int J Med Inform*. 2014;83(5):368–375.

20. Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, Lee R, et al. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif Intell.* 2023;316:103839.
21. Cheah PY, Parker M. Consent and assent in paediatric research in low-income settings. *BMC Med Ethics.* 2014;15:22.
22. Mirabella AM, Berson IR, Berson MJ. Empowering voices: Implementing ethical practices for young children's assent in digital research. *Educ Sci.* 2025;15(5):571.
23. Spriggs M. Children and bioethics: Clarifying consent and assent in medical and research settings. *Br Med Bull.* 2023;145(1):110–119.
24. Muralidharan V, Burgart A, Daneshjou R, et al. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *npj Digit Med.* 2023;6:166.
25. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int J Hum Comput Stud.* 2023;169:102941.
26. Panigutti C, Beretta A, Giannotti F, Pedreschi D. Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* ACM; 2022:1–21.
27. Hsu LM, Field R. Interrater agreement measures: Comments on kappa κ , Cohen's kappa, Scott's π , and Aickin's α . *Underst Stat.* 2003;2(3):205–219.
28. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Med Res Methodol.* 2013;13:61.
29. Shulha M, Hovdebo J, D'Souza V, Thibault F, Harmouche R. Integrating explainable machine learning in clinical decision support systems: Study involving a modified design thinking approach. *JMIR Form Res.* 2024;8:e50475.
30. Lawson DO, Mellor K, Eddy S, Lee C, Kim KH, Kim K, et al. Pilot and feasibility studies in rehabilitation research: A review and educational primer for the physiatrist researcher. *Am J Phys Med Rehabil.* 2022;101(4):372–383.
31. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. *BMC Med.* 2019;17:230.
32. Brierley J, Verhagen E; European Academy of Paediatrics Ethics Strategic Advisory Group. European Academy of Paediatrics position paper: A response to the 2024 Declaration of Helsinki update. *Acta Paediatr.* 2025;114(11):2827–2829.
33. Hein IM, De Vries MC, Troost PW, Meynen G, Van Goudoever JB, Lindauer RJL. Informed consent instead of assent is appropriate in children from the age of twelve: Policy implications of new findings on children's competence to consent to clinical research. *BMC Med Ethics.* 2015;16:76.
34. Wild CEK, Rawiri NT, Taiapa K, Anderson YC. In safe hands: Child health data storage, linkage and consent for use. *Health Promot Int.* 2023;38(6):daad159.

35. Joseph J. Algorithmic bias in public health AI: A silent threat to equity in low-resource settings. *Front Public Health*. 2025;13:1643180.
36. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence–based clinical decision support systems among health care workers: Systematic review. *J Med Internet Res*. 2025;27:e69678.
37. Rezaeian O, Bayrak AE, Asan O. Explainability and AI confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care. *Int J Hum Comput Interact*. 2025. doi:10.1080/10447318.2025.2539458.