

Toward an Explainable Digital Biomarker for Dyslexia: A Five-Phase Translational Research Program

Carrie VanZant, PhD

Neuro Navigation, Inc., Wisconsin, United States

Author Note

Correspondence concerning this article should be addressed to Carrie VanZant, PhD, Executive Director, Neuro Navigation, Inc. Email: carrievanzant@neuronavigation.org. This paper synthesizes a five-phase translational research program (Discover, Organize, Prioritize, Develop, Validate) directed toward the development of an Explainable Digital Biomarker for Dyslexia (EDBD). The program was conducted by Neuro Navigation, Inc., a corporation registered in Wisconsin in January 2025, in service of its founding mission to identify neurodivergent learners earlier and support them with tools grounded in both lived experience and rigorous science.

Running head: EXPLAINABLE DIGITAL BIOMARKER FOR DYSLEXIA

Abstract

Dyslexia affects approximately 15% to 20% of the school-age population in the United States, yet formal identification is frequently delayed until a child has already experienced years of reading failure, a pattern commonly described as the "wait-to-fail" model of identification. This delay carries measurable costs to literacy, self-concept, and long-term educational and health trajectories, and it disproportionately affects families with limited access to comprehensive psychoeducational evaluation. This paper reports a complete, five-phase translational research program undertaken by Neuro Navigation, Inc. toward the development of an Explainable Digital Biomarker for Dyslexia (EDBD): a transparent, evidence-based, low-cost instrument designed to estimate dyslexia probability within a fifteen- to twenty-minute digital session and to help schools, clinicians, and families act sooner.

Phase 1 (Discover) conducted a PRISMA-informed systematic review of four decades (1986–2026) of peer-reviewed literature, identifying validated dyslexia markers across nine domains: phonological processing, rapid automatized naming, orthographic processing, working memory, morphological awareness, neurobiological markers, oculomotor and eye-movement markers, genetic markers, and digital and gamified assessment. Phonological awareness and rapid automatized naming (RAN) emerged as the strongest and most reproducible markers, with strong neurobiological and genetic evidence converging on a well-validated phenotype. Phase 2 (Organize) reorganized the resulting 24-marker catalog into a structured, field-coded evidence atlas, coding each marker on ten standardized fields: definition, evidence strength, age applicability, languages studied, orthographic transparency, measurement method, digital feasibility, psychometric quality, replication status, and key citations. Phase 3 (Prioritize) applied a seven-criterion weighted rubric (evidence strength, replication consistency, effect size,

feasibility, digital measurability, interpretability, and clinical usefulness) to rank all 24 markers and identified a four-marker minimum set (phonemic awareness, RAN, orthographic-choice tasks, and serial-order verbal working memory) estimated to require 12 to 15 minutes of digital administration. Phase 4 (Develop) translated this marker specification into a complete, explainable system architecture: an additive, intrinsically interpretable scoring model whose probability estimate can be decomposed marker by marker, paired with statistical uncertainty intervals and plain-language output designed for both clinicians and families. Phase 5 (Validate) specifies a prospective, school-based pilot study protocol following STARD and STARD-AI reporting standards, designed to test feasibility, generate preliminary diagnostic accuracy estimates, and evaluate the EDBD's explanation interface in real classrooms.

Together, these five phases represent the first complete, evidence-grounded path from what the science tells us about dyslexia to a tool designed to act on it in real schools. This research culminates in a call for school-based piloting, because the conviction driving this work is both simple and urgent: no child should spend years struggling to read before anyone can tell them why. Children like Kelly VanZant, Neuro Navigation's founder, who was not identified with dyslexia and ADHD until age nine, deserve a system that finds them sooner.

Keywords: dyslexia, digital biomarker, explainable artificial intelligence, early identification, translational research, phonological awareness, rapid automatized naming, implementation science, public health

Introduction

Dyslexia is the most common of the specific learning disabilities, affecting approximately 15% to 20% of the U.S. school-age population. In any average classroom of 25 students, four or five children are likely affected (International Dyslexia Association, 2025). Despite four decades of rigorous scientific inquiry into the cognitive and neurobiological architecture of dyslexia, the practical experience of most affected children remains largely unchanged. Reading difficulty is noticed informally by a parent or teacher, often dismissed or attributed to immaturity or lack of effort, and not formally evaluated until the child has already fallen substantially behind grade-level expectations. This pattern, widely termed the "wait-to-fail" model, persists even where Response to Intervention (RTI) frameworks were designed to prevent it, because RTI itself frequently requires documented failure at one instructional tier before a student is escalated to more intensive support.

The consequences of delayed identification are not merely academic. Developmental neuroscience is clear: the brain's capacity for the kind of learning that underlies phonological and orthographic skill is greatest in early childhood and declines measurably after the primary grades. Interventions delivered in kindergarten and first grade produce larger and more durable gains than functionally identical interventions delivered in third grade and beyond. Every year a child waits is a year that becomes harder to recover. Delayed identification also compounds secondary consequences. Repeated reading failure is associated with diminished academic self-concept, elevated anxiety, and disengagement from schooling, effects that often persist into adulthood long after any gains in decoding have been achieved. From a public health perspective, delayed identification of dyslexia is not a narrow pedagogical inconvenience. It is a

preventable driver of educational, occupational, and mental health disparities, one that falls hardest on the families who can least afford to wait.

The current identification pathway for children with dyslexia in the United States operates primarily through the Child Find mandate of the Individuals with Disabilities Education Act (IDEA, 2004), which requires states and local education agencies to identify, locate, and evaluate all children who may have a disability and require special education services. Despite this mandate, identification remains largely reactive: referrals are typically initiated after sustained academic failure is observed by a teacher or parent, resulting in an average age at diagnosis of 9 to 10 years, well after the critical window for early reading intervention has narrowed (Shaywitz et al., 2008; Lyon et al., 2003). This "wait-to-fail" dynamic is compounded by significant inequities in evaluation access across school districts stratified by resource level, race, and socioeconomic status (Morgan et al., 2015; Shifrer et al., 2011). The consequence is that children who would benefit most from early, structured literacy intervention accumulate 2 to 4 years of compounding academic difficulty before they receive a formal determination. A proactive, equitable screening instrument capable of estimating dyslexia probability before observable failure, and before formal evaluation, could compress this identification gap and improve downstream outcomes at a population level.

Neuro Navigation, Inc. is a Wisconsin-registered corporation founded in January 2025 to provide resources, tools, and services nationally for neurodivergent learners and their families, with a particular focus on dyslexia and ADHD, conditions that together touch approximately 15% of the total U.S. student population. The company's founder and creative director, Kelly VanZant, was identified with ADHD, dyslexia, dyscalculia, and auditory processing delays at age nine, well past the point at which current evidence indicates her reading difficulty was almost

certainly already detectable. Neuro Navigation's mission is grounded in Kelly's lived experience of a system that too often waits for a child to fail before it acts, and in a single conviction: no child should spend years struggling with reading before anyone can tell them why.

Toward that end, Neuro Navigation undertook a five-phase translational research program directed at the development of an Explainable Digital Biomarker for Dyslexia (EDBD): an assessment that is evidence-based, transparent in its reasoning, inexpensive to administer, completable in 15 to 20 minutes in a digital format, and capable of producing a dyslexia probability estimate that helps schools, clinicians, and families act sooner rather than waiting for a child to fail. The five phases (Discover, Organize, Prioritize, Develop, and Validate) are reported as a complete program in this paper, with each phase building directly on the deliverable of the preceding one.

The remainder of this paper proceeds as follows. The Method section describes the research design and procedures for each of the five phases. The Results section presents the key findings and deliverables of each phase sequentially. The Discussion synthesizes findings across phases, addresses limitations, and situates this work within the broader literature on responsible AI in education and public health. The Conclusion returns to the founding mission and the pathway from completed specification to empirical validation.

Method

This paper reports five sequential, linked research phases, each with its own design and procedures. Phases 1 through 4 are evidence synthesis and specification studies; no new data from children were collected as part of any of these phases. Phase 5 is a prospective study

protocol; it describes planned procedures rather than completed data collection. Each phase is described in the order in which it was conducted.

Phase 1 — Discover: Systematic Review

Phase 1 conducted a PRISMA-informed systematic review of the peer-reviewed dyslexia biomarker literature published between 1986 and 2026. The 40-year window was selected to capture the field from the earliest rapid-naming-speed and phonological-deficit research programs through the most current digital, gamified, and explainable-AI screening literature. Literature was identified through a structured search of peer-reviewed academic databases spanning coverage equivalent to PubMed, Scopus, and Semantic Scholar, supplemented by targeted verification of foundational sources. Search terms combined "dyslexia" and "reading disability" with domain-specific terms spanning phonological awareness, rapid automatized naming, orthographic processing, working memory, morphological awareness, functional MRI, EEG, eye-tracking, genetic markers, and digital or gamified assessment.

Studies were eligible if they were peer-reviewed, involved human participants, addressed dyslexia as a primary construct, and were reported in English. Meta-analyses, systematic reviews, and large-sample primary studies were preferentially retained. Studies were excluded if they could not be bibliographically verified or were available only as preprints, conference abstracts, or gray literature. Evidence was synthesized across nine marker domains (phonological processing, rapid automatized naming, orthographic processing, working memory, morphological awareness, neurobiological markers, oculomotor and eye-movement markers, genetic markers, and digital and gamified assessment), with attention to three evaluative

dimensions: evidence strength, reproducibility, and digital measurability. Fifty-seven peer-reviewed sources meeting eligibility criteria were synthesized.

Phase 2 — Organize: Evidence Atlas

Phase 2 took the Phase 1 marker catalog and reorganized it into a structured, queryable evidence atlas. Twenty-four markers across eleven domains were coded on ten standardized fields: definition, evidence strength, age applicability, language(s) studied, orthographic transparency, measurement method, digital feasibility, psychometric quality, replication status, and key citations. Classification drew on the Phase 1 catalog and a targeted supplementary search addressing fields the original review left thin, most notably cross-linguistic orthographic transparency and reliability and validity evidence for individual instruments. A companion structured database was produced alongside the atlas.

Phase 3 — Prioritize: Marker Prioritization

Phase 3 developed and applied a seven-criterion weighted rubric to rank all 24 atlas markers and identify a minimum marker set for prototype development. Criteria were evidence strength (20%), replication consistency (15%), effect size (15%), feasibility (15%), digital measurability (10%), interpretability (10%), and clinical usefulness (15%). Each criterion was operationally defined with explicit conversion rules translating the atlas's qualitative fields into a 0–5 numeric scale. Scores were computed before results were inspected, and no weights were adjusted post-hoc. A supplementary literature search addressed three questions Phase 2 had not needed to answer: the statistical independence of phonological awareness and RAN, the relationship between screening-battery length and accuracy, and the operational definition of interpretability in the explainable-AI literature.

Phase 4 — Develop: Framework Specification

Phase 4 translated the Phase 3 marker specification into a complete system architecture for the EDBD. No children were assessed during this phase; the deliverable is a fully justified, testable design specification. Framework development required four linked activities: translating Phase 3's specification and the program's standing constraints into formal architecture requirements; a targeted literature search across explainable AI methods, clinical prediction modeling standards, and existing digital dyslexia-screening tools; selecting and justifying a single architecture against those requirements; and specifying each additional framework component (weighting, uncertainty estimation, visualization, interpretability safeguards, and reasoning transparency) to the same evidence-traceability standard. Every design choice is defended by a citation to a specific finding in the relevant literature.

Phase 5 — Validate: Pilot Study Protocol

Phase 5 specifies a prospective, single-arm, two-stage pilot study protocol following the Standards for Reporting of Diagnostic Accuracy Studies (STARD 2015) and its AI-specific extension (STARD-AI). The study is explicitly designed as a pilot rather than a definitive validation trial, using pre-specified feasibility progression criteria to determine readiness for a subsequent, fully powered multi-site trial. The protocol addresses three co-equal objectives: operational feasibility, preliminary diagnostic accuracy, and explainability and human-factors performance. Target enrollment is approximately 200 children in grades 1–3 for universal EDBD screening, with a stratified sub-sample of approximately 60 children receiving a blinded comprehensive reference-standard evaluation. All planned analyses are descriptive and exploratory; no p-values or significance tests will be applied to diagnostic accuracy estimates.

Results

Phase 1 — Discover: The Dyslexia Marker Landscape

Fifty-seven peer-reviewed sources met eligibility criteria. Evidence was synthesized across nine domains: phonological processing, rapid automatized naming, orthographic processing, working memory, morphological awareness, neurobiological markers, oculomotor and eye-movement markers, genetic markers, and digital and gamified assessment, moving from the most extensively replicated cognitive-linguistic markers toward emerging neurobiological and digital ones. Key findings by domain are reported below, culminating in a Marker Catalog (Table 1) that informed Phase 2.

Phonological Processing

Phonological processing, and phonemic awareness in particular, is the most extensively validated cognitive marker of dyslexia identified in this review. The largest meta-analysis in this domain (Melby-Lervåg et al., 2012; 235 studies, 995 effect sizes) reported a large deficit in phonemic awareness relative to age-matched typically developing peers (pooled $g \approx -1.37$). Longitudinal and cross-linguistic studies converge on this finding: phonological awareness measured as early as 3.5 years of age, combined with familial risk status and letter knowledge, produced an area under the receiver operating characteristic curve above .80 for later reading disability (Puolakanaho et al., 2007). Phonological awareness is readily adaptable to brief, audio-based digital tasks, making it a strong candidate for the EDBD.

Rapid Automatized Naming

Rapid Automatized Naming (RAN), which measures the speed with which an individual names familiar visual stimuli, is the second most consistently validated marker. Wolf and Bowers

(1999) formalized the double-deficit hypothesis, proposing that phonological-awareness and naming-speed deficits are separable sources of reading dysfunction, with children showing both deficits exhibiting more severe and pervasive impairment than either single-deficit group.

Cross-linguistic replication is a particular strength: the double-deficit pattern has been demonstrated in English, German, Finnish, Spanish, Greek, and Chinese samples (Torppa et al., 2012, 2013; Li et al., 2022). RAN tasks require only a stimulus display and a timed response, making them straightforwardly adaptable to digital administration.

Orthographic Processing, Working Memory, and Morphological Awareness

Orthographic processing shows a moderate-to-strong evidentiary base sensitive to orthographic depth, with larger deficits documented in opaque orthographies such as English (Carioti et al., 2021). Verbal and phonological working memory deficits are consistently reported in dyslexia, with serial-order short-term memory impairment demonstrating particular specificity (Majerus et al., 2016; Cowan et al., 2017). Morphological awareness shows a large effect size against chronological-age controls ($g = 1.11$) but a non-significant comparison against reading-level-matched controls, indicating this marker tracks general reading development rather than acting as a primary, dyslexia-specific signal (Georgiou et al., 2022).

Neurobiological, Genetic, and Digital Markers

Structural and functional neuroimaging provide the most extensive biological validation, with four independent meta-analyses converging on hypoactivation in left occipito-temporal and temporo-parietal regions during reading tasks in children with dyslexia (Maisog et al., 2008; Paulesu et al., 2014; Richlan et al., 2013; Linkersdörfer et al., 2012). These findings confirm dyslexia as a neurobiologically real phenotype, though MRI and fMRI are incompatible with the

EDBD's low-cost, brief, digital format. Genetic markers establish substantial heritability (40%–60% in twin studies) but show frequent non-replication at the individual candidate-gene level (Becker et al., 2013), supporting a highly polygenic architecture not yet amenable to same-session digital assessment. The digital and gamified assessment domain shows the most direct relevance to the EDBD's design goals: published classification accuracies ranging from 67% to 96%, with cross-linguistic validation in some tools (Rello et al., 2019; Rauschenberger et al., 2025; Zugarramurdi et al., 2022). Methodological heterogeneity and overfitting concerns temper these findings.

Table 1

Summary Marker Catalog: Evidence Strength, Reproducibility, and Digital Measurability

Domain	Marker	Evidence Strength	Reproducibility	Digitally Measurable
Phonological processing	Phonemic awareness	Very Strong (g ≈ -1.37)	High — 40-year cross-linguistic replication	Yes — audio-based tasks, automated scoring
Rapid naming	RAN composite	Strong — additive with PA (double-deficit)	High — replicated across 6+ orthographies	Yes — timed digital naming, scalable

Domain	Marker	Evidence Strength	Reproducibility	Digitally Measurable
Orthographic processing	Orthographic-choice tasks	Moderate–Strong (d = 1.7–2.0 in adults)	Moderate — depth-moderated	Yes — forced-choice paradigm, screen-based
Working memory	Serial-order verbal WM	Moderate — strongest digital reliability	Moderate — heterogeneous profiles	Yes — digital span tasks, $\alpha = .88$
Morphological awareness	Inflectional/derivational tasks	Strong vs. age-matched (g = 1.11); null vs. reading-level	Moderate — large cross-linguistic base	Yes — sentence completion tasks
Neurobiological (structural)	Gray-matter volume — L temporo-parietal	Moderate–Strong — convergent ALE meta-analyses	Moderate	No — requires MRI
Neurobiological (functional)	Hypoactivation — L occipito-temporal	Strong — most replicated neuroimaging finding	High across cross-linguistic fMRI meta-analyses	No — requires fMRI
Eye movements	Fixation duration/count	Mixed — inconsistent	Low–Moderate	Emerging — webcam

Domain	Marker	Evidence Strength	Reproducibility	Digitally Measurable
		group differences		eye-tracking feasible
Genetic	DCDC2, KIAA0319 variants	Moderate — 20–25% SNP-based heritability	Low–Moderate — frequent non-replication	No — not a same-session digital measure
Digital/gamified	Game-based phonological + ML	Moderate–Strong and emerging (67%–96% accuracy)	Moderate — validation heterogeneous	Yes — by definition

Phase 2 — Organize: The Dyslexia Evidence Atlas

Phase 2 reorganized the Phase 1 marker catalog into a structured, field-coded evidence atlas of 24 markers across 11 domains. Each marker was coded on 10 standardized fields. The atlas confirmed that phonemic awareness and RAN are the only markers that combine strong evidence, high cross-linguistic replication, documented psychometric adequacy (test-retest reliability $\geq .84$), and full digital measurability. A companion structured database (the Dyslexia Evidence Atlas Database) was produced as a queryable resource for subsequent phases.

The atlas identified several critical distinctions that a narrative review alone could not capture. Orthographic-choice tasks showed strong effect sizes ($d > 1.7$ in opaque orthographies) but an orthographic-depth moderation requiring explicit calibration before cross-linguistic

deployment. Serial-order verbal working memory showed the strongest digital reliability profile of any working-memory marker (Cronbach's $\alpha = .88$; ICC up to .99), despite more heterogeneous theoretical accounts of the underlying deficit. Morphological awareness and eye-tracking markers showed real and theoretically important signals but incomplete psychometric documentation for a founding-tier instrument. Neurobiological and genetic markers were documented as important convergent validation evidence rather than direct instrument inputs, given equipment and scalability barriers.

Phase 3 — Prioritize: The Four-Marker Minimum Set

Phase 3 applied a seven-criterion weighted rubric to all 24 atlas markers. The seven criteria were evidence strength (20%), replication consistency (15%), effect size (15%), feasibility (15%), digital measurability (10%), interpretability (10%), and clinical usefulness (15%). Phonemic awareness and RAN each reached the maximum composite score (100.0%), reflecting strong performance across all seven criteria simultaneously. Orthographic-choice tasks ranked third (89.0%), driven by a very large effect size offset by thinner formal reliability documentation for the specific forced-choice paradigm. Serial-order verbal working memory ranked fourth (80.0%), adding a distinct memory-based axis with the strongest digital reliability profile of any working-memory marker.

Table 2

Proposed Four-Marker Minimum Set for the EDBD

Construct Axis	Marker	Compos ite %	Est. Time	Rationale
Phonological	Phonemic awareness (deletion, elision, blending)	100.0	~4 min	Anchor marker; 40-yr cross-ling. replication; standardized reliability $\geq .80$
Naming speed	RAN — composite (letters/digits/objects/c olors)	100.0	~3 min	Independent axis from PA per double-deficit literature; test-retest .84–.92
Orthographic	Orthographic-choice/k knowledge tasks	89.0	~4 min	Captures print-specific representation not indexed by PA or RAN; depth-calibration required
Working memory	Verbal working memory — serial order	80.0	~4 min	Distinct memory axis; $\alpha = .88$, ICC up to .99; strongest digital reliability in atlas

Four additional Tier 1 markers were identified as redundant with the minimum set rather than as independent additions: the RAN alphanumeric subtype (a stimulus-type component of the RAN composite already included), the double-deficit profile (an arithmetic derivative of phonemic awareness and RAN, recommended as a downstream severity flag), and both verbal short-term memory and central-executive working memory (both occupying the same construct axis as serial-order working memory but with weaker reliability documentation). Four Tier 2 markers (web-based gamified screening, morphological awareness, visual attention span, and eye-tracking) were classified as complementary to the minimum set but not yet sufficiently

reliability-documented for a founding-tier instrument, and are recommended for staged development in future iterations.

An orthographic-depth calibration requirement was identified as a key open parameter: the orthographic-choice task's effect size is moderated by orthographic transparency, with larger deficits documented in opaque orthographies such as English and smaller deficits in more transparent ones. This parameter, labeled λ_{ortho} , must be empirically resolved before national or international deployment in orthographies differing materially from English.

Phase 4 — Develop: The EDBD Framework Specification

Phase 4 translated the four-marker specification into a complete, testable system architecture. The EDBD Framework Specification comprises five integrated components: a constrained additive scoring architecture; an evidence-informed initial marker weighting scheme; a two-layer confidence and uncertainty estimation approach; a dual-audience visualization design; and a transparent reasoning mechanism. The overarching architecture is summarized in Table 3.

Table 3

EDBD Five-Layer Architecture Overview

Layer	Function	Key Elements
1. Input	Administer the four Phase 3 markers plus brief intake	Phonemic awareness; RAN composite; orthographic-choice tasks; serial-order WM; demographic/orthography intake

Layer	Function	Key Elements
2. Feature	Convert raw task output to normalized, age-referenced scores	Per-marker normalization against age-based norms; data-quality checks (completion, timing, audio signal)
3. Scoring	Combine marker scores into a single log-odds estimate	Additive model: $\text{logit}(p) = \beta_0 + f_1(\text{PA}) + f_2(\text{RAN}) + f_3(\text{Ortho}) + f_4(\text{WM}) + f_5(\text{PA} \times \text{RAN})$
4. Confidence	Quantify uncertainty in the estimate	Conformal prediction interval (statistical); completion-quality flag (data-quality)
5. Output	Render estimate for two audiences	Family/educator icon array + plain-language narrative; clinician bar chart + normed reference panel

Architecture Selection and Justification

This paper selected a constrained additive architecture over post-hoc explanation of an opaque model, for three reasons grounded in the reviewed literature. First, Rudin (2019) argues that post-hoc explanation methods applied to black-box models cannot guarantee fidelity to the model's actual reasoning, and that for tabular problems with a moderate number of features, precisely the EDBD's situation, interpretable models are frequently competitive with black-box alternatives on accuracy. Second, Ghassemi et al. (2021) extend this concern specifically to healthcare, arguing that current explainability methods do not reliably tell a clinician whether an individual prediction was reached for a clinically valid reason. Third, Caruana et al. (2015)

demonstrated that generalized additive models with pairwise interactions (GA2Ms) achieved accuracy competitive with state-of-the-art black-box alternatives on real healthcare prediction problems while remaining fully inspectable term by term.

The EDBD's core scoring function is formally specified as: $\text{logit}(p) = \beta_0 + f_1(\text{PA}) + f_2(\text{RAN}) + f_3(\text{Ortho}) + f_4(\text{WM})$, where p is the estimated probability of a dyslexia-consistent profile and each f_i is a monotonic, piecewise-linear shape function relating that marker's normalized score to its contribution to the overall log-odds estimate. A single pairwise interaction term, $f_5(\text{PA} \times \text{RAN})$, captures the double-deficit pattern identified in Phase 3 as a downstream derived severity flag.

Marker Weighting

Initial marker weights were derived directly from Phase 3's composite rubric scores, normalized to sum to 100%: phonemic awareness ~27%, RAN composite ~27%, orthographic-choice tasks ~24%, and serial-order working memory ~22%. These are not fixed regression coefficients but starting values for estimation in Phase 5, following standard practice in clinical prediction model development (Steyerberg, 2019). Phase 5 will re-estimate each shape function and associated weight via penalized additive logistic regression on labeled pilot data, using Phase 3-derived weights as an informative prior.

Confidence Estimation

The framework treats uncertainty as having two distinct sources. Statistical uncertainty is addressed through a conformal prediction wrapper, producing distribution-free, finite-sample coverage guarantees without requiring strong model assumptions — a property especially important at an early pilot stage. Data-quality uncertainty is addressed through a session-level

completion-quality index, computed from within-session metrics including item completion rate, response-time variability, and audio signal quality. Administrations are flagged as high, moderate, or low quality; low-quality flags suppress the numeric probability in favor of "inconclusive; repeat administration recommended." Both sources of uncertainty are always surfaced together in the clinician-facing view.

Visualization and Transparency

The family- and educator-facing tier presents probability as an icon array — a 10×10 grid in which shaded icons represent estimated probability as a countable frequency, documented to improve comprehension of probabilistic information among individuals with limited numeracy (Galesic et al., 2009). The clinician-facing tier presents a horizontal bar chart of each marker's additive log-odds contribution, annotated with the marker's raw normalized score, its Phase 3 evidence tier, and its share of the conformal interval. Both tiers are generated from the identical underlying decomposition, eliminating the "two-truths problem" in which simplified and technical explanations diverge.

The transparent reasoning mechanism populates a plain-language sentence template naming the specific constructs driving any given estimate — for example: "This estimate reflects lower accuracy on tasks that measure the ability to isolate and manipulate individual speech sounds (phonemic awareness) and slower performance on rapid-naming tasks." When both PA and RAN contributions exceed a specified threshold, a double-deficit severity note is automatically appended. When λ_{ortho} differs from 1.0, the reasoning output explicitly states and reports the locale adjustment.

Interpretability Safeguards

Three considerations shaped the framework's automation-bias protections: health literacy, actionability, and documented risk of automation bias (Goddard et al., 2012). Safeguards include fixed non-removable non-diagnostic disclaimers, always-visible raw per-marker evidence alongside any probability, completion-quality suppression of false-precision outputs, and explicit positioning of the EDBD within existing RTI and evaluation workflows rather than as a replacement for clinical judgment. Output tiers map directly to actionable next steps: low probability (< 15%) → continue universal screening; moderate (15%–40%) → initiate or intensify Tier 2 intervention; elevated (> 40%) → refer for comprehensive evaluation. These thresholds are explicitly provisional pending Phase 5 calibration.

Phase 5 — Validate: Pilot Study Protocol

Phase 5 specifies a prospective, school-based pilot study designed to generate the first empirical evidence for the EDBD specified in Phase 4. The protocol structure follows STARD 2015 and STARD-AI, with pre-specified feasibility progression criteria using a three-zone (red/amber/green) traffic-light framework (Table 4) to make the decision to proceed to a full validation trial transparent and falsifiable.

Table 4

Pre-Specified Feasibility Progression Criteria (Phase 5)

Feasibility Domain	Outcome Measure	Green (Proceed)	Amber (Modify)	Red (Reconsider)
Recruitment	% of eligible families providing consent	$\geq 60\%$	40–59%	$< 40\%$
Completion	% of consented children completing full battery	$\geq 90\%$	75–89%	$< 75\%$
Administration time	Mean wall-clock session duration	≤ 30 min	31–40 min	> 40 min
Technical reliability	% of sessions without device/software failure	$\geq 95\%$	85–94%	$< 85\%$

Target enrollment is approximately 200 children in grades 1–3 from at least two partnering elementary schools, selected to reflect variation in socioeconomic composition and prior exposure to universal literacy screening. A nested sub-sample of approximately 60 children — all screen-positive children and a randomly sampled group of screen-negative children in a 1:1 to 2:1 ratio — will undergo a comprehensive, blinded reference-standard psychoeducational evaluation. All accuracy estimates will be reported with Clopper-Pearson confidence intervals appropriate to pilot-scale samples and explicitly framed as hypothesis-generating rather than confirmatory.

A separate nested explainability and trust sub-study will enroll approximately 15–20 educators and clinicians to evaluate the EDBD's Layer 2 explanation interface. Participants will review de-identified case vignettes under two conditions — full explanation versus bare probability score — in counterbalanced order, rating clinical judgment, confidence, and agreement with the EDBD's recommendation (summarized via Cohen's kappa), followed by the

System Usability Scale and a validated clinical-AI trust instrument. This sub-study uses simulated vignettes rather than live participant cases to allow explainability evaluation before establishing diagnostic performance, and to protect children's data from a system whose accuracy has not yet been established.

Ethical procedures follow the ACCEPT-AI framework's recommendations for responsible inclusion of pediatric data in AI/ML research, a dual-layer permission process (written parental informed consent followed by separate, age-appropriate child assent), and alignment with the 2024 Declaration of Helsinki's emphasis on ongoing, iterative consent. No individual EDBD result will be shared with a family or teacher without accompanying interpretive guidance from a study clinician, and screening results will not be used on their own to make any special-education eligibility determination.

Discussion

This five-phase program represents the first complete, evidence-grounded path from what the science tells us about dyslexia to a tool that can act on it in real schools. The sequencing was deliberate: evidence synthesis and architectural specification first, followed by staged, transparently planned empirical testing. This reflects a principle the authors hold firmly about how AI tools for children should be built. Not by claiming novelty and validated accuracy at the same time, but by earning the right to enter a classroom through rigorous, documented preparation.

Contribution to the Field

Four aspects of this program's contribution merit specific discussion. First, the EDBD is the first proposed dyslexia screening instrument whose marker selection, weighting, and scoring architecture are each individually traceable to specific, citable evidence rather than to convenience or convention. Every marker in the minimum set was admitted for a documented reason; every marker excluded was excluded for an equally documented reason. This traceability is not merely a procedural virtue — it is what makes the instrument revisable as the evidence base evolves, without requiring a complete restart.

Second, the architecture's commitment to intrinsic interpretability over post-hoc explanation of a black-box model addresses a gap documented across both the general explainable AI literature (Rudin, 2019; Ghassemi et al., 2021) and the dyslexia-specific digital screening literature (Pamungkas et al., 2025; Robaa et al., 2024). Existing digital dyslexia screening tools achieve strong classification accuracy but document that educators and clinicians frequently do not trust outputs they cannot interpret. The EDBD's additive architecture is designed to be the explanation rather than to carry a separate explanation layered on top of it.

Third, the conformal prediction approach to uncertainty quantification directly addresses a gap Ghassemi et al. (2021) identify as central to trustworthiness in medical AI: most deployed clinical AI systems report only a point estimate, providing no mechanism for distinguishing high-confidence from low-confidence individual predictions. The completion-quality flag adds a further layer of per-session data quality transparency that statistical uncertainty estimation alone cannot capture.

Fourth, and most directly relevant to the practical problem this program was built to solve, the program's explicit focus on digital feasibility and plain-language explainability throughout — from Phase 1's digital-measurability ratings through Phase 4's icon-array

visualization and plain-language reasoning templates — is what distinguishes the EDBD from a research-grade assessment instrument. The goal is not to describe what could be measured in a laboratory; it is to specify what can be measured, explained, and acted on in a real elementary school by a classroom teacher or a parent who is not a reading specialist.

Implications for Equity

The wait-to-fail pattern this program exists to disrupt is not equally distributed. Access to comprehensive psychoeducational evaluation, the current gold standard for a dyslexia determination, costs \$400 to \$800 per child in the private sector and can require months-long waits for a licensed evaluator (National Center for Learning Disabilities, 2014). Children whose families can afford private evaluation are identified sooner. Children whose families rely on public-sector evaluation are identified later, if at all. A brief, digitally administered, low-cost instrument whose output is explainable to a parent without clinical training does not close this equity gap by itself, but it is a specific, addressable lever that this program treats as foundational rather than optional. A design target of 15 to 20 minutes, consumer digital hardware, and output any non-specialist adult can understand was chosen deliberately, because a more expensive, more complex, or more technically demanding instrument would simply replicate the access barriers this work is meant to dismantle.

This equity framing is also why Phase 5's protocol explicitly diversifies recruitment across schools with different socioeconomic compositions and prior screening exposure, rather than piloting in a single high-resource school where completion rates and accuracy estimates would be most favorable. Neuro Navigation's mission is to serve the full national population of

neurodivergent learners, not a demographically narrow subset of it, and the Phase 5 design reflects that commitment.

It is important to situate the EDBD within, rather than apart from, the existing legal framework for disability identification. The EDBD produces a probabilistic screening score — not a diagnosis and not an eligibility determination under IDEA. It is not designed to replace, and cannot legally substitute for, the multidisciplinary evaluation required by Child Find. Rather, the EDBD is positioned as a proactive first stage in a two-stage identification pipeline: universal digital screening generates an early, explainable probability estimate that provides documented clinical signal for initiating a timely Child Find referral. The formal evaluation then proceeds through established channels, with the EDBD output serving as one data point among those considered by the multidisciplinary team. This architecture preserves the legal protections and procedural safeguards of IDEA while addressing the reactive posture that delays identification. Importantly, because the EDBD is designed for universal administration rather than selective referral, it has the potential to reduce the documented disparities in which children are identified — disparities shaped by race, language, and socioeconomic access to advocacy (Morgan et al., 2015) — by decoupling early screening from the subjective referral process.

Limitations

Several limitations should temper how this program is used. First, Phases 1 through 4 are synthesis and specification studies; no data have been fit to the proposed architecture, and no numeric element of the EDBD Framework Specification — initial marker weights, probability tier thresholds, conformal coverage guarantee, or the orthographic-depth calibration coefficient — has yet been empirically validated. All are provisional pending Phase 5. Second, the Phase 1

evidence base is weighted toward English and a limited set of well-studied European languages; markers that would perform differently in logographic or abjad writing systems are not yet adequately covered. Third, the Phase 3 rubric was applied by a single research team rather than through a multi-rater process with a formal agreement statistic, which cannot rule out systematic scoring bias. Fourth, the framework's interpretability and visualization design choices are grounded in the published health-risk-communication and clinical-XAI literatures rather than in original user testing with the parents, teachers, and clinicians who would be the EDBD's actual users; Phase 5's nested usability sub-study is the first opportunity to close that gap. Fifth, the anticipated pilot sample size is explicitly not powered for confirmatory diagnostic accuracy estimation; the resulting sensitivity, specificity, and AUC figures will be hypothesis-generating inputs to the design of a subsequent, fully powered trial rather than established clinical accuracy claims.

Future Directions

If Phase 5's feasibility and preliminary accuracy findings support progression, the resulting fully powered, multi-site diagnostic accuracy trial will be the first study in this program to generate genuinely new data about children rather than to synthesize, prioritize, or specify based on existing publications. That threshold — moving from evidence synthesis to empirical contact with the population the instrument is meant to serve — has been treated throughout this program as consequential enough to warrant the full ethical and methodological infrastructure Phase 5 describes.

Beyond its direct output, this program is designed to contribute to two literatures that Dr. Carrie VanZant's research agenda 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 contexts; comparatively little evaluates whether explanation interfaces function similarly when the end users are classroom teachers and school psychologists rather than physicians, and when the subjects of prediction are children rather than adult patients. The Phase 5 nested trust sub-study is designed to begin filling that gap. Future publication targets include: a responsible AI framework for supporting students with dyslexia, implementation science for AI in education, and digital health equity for neurodivergent populations.

Conclusion

Four decades of peer-reviewed research have produced a rich, convergent, and increasingly digitally actionable evidence base for identifying dyslexia. This program has organized, prioritized, and translated that evidence into a complete, testable system specification: a four-marker minimum instrument (phonemic awareness, RAN, orthographic-choice tasks, and serial-order verbal working memory) integrated through an inherently interpretable additive architecture, with conformal uncertainty quantification, dual-audience visualization, automation-bias safeguards, and a plain-language reasoning mechanism that decomposes any probability estimate into marker-level statements a parent can follow as readily as a specialist.

What this program has accomplished is something foundational: a fully specified, literature-grounded architecture where every design choice is traceable to evidence rather than convenience, and where every open question is named explicitly rather than left to assumption. Phase 5 will now test whether that architecture performs as designed when it leaves the specification and enters an actual classroom. The founding conviction of this work is that no

child should spend years struggling with reading before anyone can tell them why. That conviction does not by itself specify which four markers belong in a 15-minute digital assessment. These five phases are where that conviction and that specification meet. Phase 5 is where they will be put to the test, and where the work of making the promise real finally begins.

References

- Becker, J., Czamara, D., Scerri, T. S., Ramus, F., Csépe, V., Talcott, J. B., Stein, J., Morris, A., Ludwig, K. U., Hoffmann, P., ... Schulte-Körne, G. (2013). Genetic analysis of dyslexia candidate genes in the European cross-linguistic NeuroDys cohort. *European Journal of Human Genetics*, 22(5), 675–680.
- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., ... Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, 351, h5527.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligent models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.
- Carioti, D., Masia, M. F., Travellini, S., & Berlingeri, M. (2021). Orthographic depth and developmental dyslexia: A meta-analytic study. *Annals of Dyslexia*, 71(3), 399–438.
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378.
- Cowan, N., Hismjatullina, A., AuBuchon, A. M., Saults, J. S., Horton, N., Leadbitter, K., & Elliott, E. (2017). Short-term memory in childhood dyslexia: Deficient serial order in multiple modalities. *Dyslexia*, 23(3), 209–233.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, 28(2), 210–216.

Georgiou, G. K., Martinez, D., Alkhalifah, S., & Torppa, M. (2022). A meta-analysis of morphological awareness deficits in developmental dyslexia. *Scientific Studies of Reading*, 26(6), 508–527.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750.

Gialluisi, A., Andlauer, T. F. M., Mirza-Schreiber, N., Moll, K., Becker, J., Hoffmann, P., ... Fisher, S. E. (2020). Genome-wide association study reveals new insights into the heritability and genetic correlates of developmental dyslexia. *Molecular Psychiatry*, 26(7), 3004–3017.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

International Dyslexia Association. (2025). Definition of dyslexia.
<https://dyslexiaida.org/definition-of-dyslexia/>

Li, M., Christ, S. L., Zhang, D., Wilson, S., Loeb, D., Ferrari, P., & Grigorenko, E. L. (2024). Phonological awareness mediates the relationship between DCDC2 and reading performance with home environment. *npj Science of Learning*, 9, 27.

Linkersdörfer, J., Lonnemann, J., Lindberg, S., Hasselhorn, M., & Fiebach, C. J. (2012). Grey matter alterations co-localize with functional abnormalities in developmental dyslexia: An ALE meta-analysis. *PLoS ONE*, 7(8), e43122.

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia*, 53(1), 1–14.

Maisog, J. M., Einbinder, E. R., Flowers, D. L., Turkeltaub, P. E., & Eden, G. F. (2008). A meta-analysis of functional neuroimaging studies of dyslexia. *Annals of the New York Academy of Sciences*, 1145(1), 237–259.

Majerus, S., & Cowan, N. (2016). The nature of verbal short-term impairment in dyslexia: The importance of serial order. *Frontiers in Psychology*, 7, 1522.

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, 138(2), 322–352.

Morgan, P. L., Farkas, G., Cook, M., Strassfeld, N. M., Hillemeier, M. M., Pun, W. H., & Schussler, D. L. (2015). Are Hispanic-, Asian-, Native-, and African-American children disproportionately underrepresented in special education? *Educational Researcher*, 44(5), 278–292.

National Center for Learning Disabilities. (2014). *The state of learning disabilities: Facts, trends, and emerging issues* (3rd ed.). NCLD.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.

Pamungkas, Y., Rangkuti, R. Y., Karim, A., & Sangsawang, T. (2025). Recent advances in artificial intelligence for dyslexia detection: A systematic review. *International Journal of Robotics and Control Systems*, 5(3), 2016–2033.

Paulesu, E., Danelli, L., & Berlingeri, M. (2014). Reading the dyslexic brain: Multiple dysfunctional routes revealed by a new meta-analysis of PET and fMRI activation studies. *Frontiers in Human Neuroscience*, 8, 830.

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H. T., Poikkeus, A.-M., ... Lyytinen, H. (2007). Very early phonological and language skills: Estimating individual risk of reading disability. *Journal of Child Psychology and Psychiatry*, 48(9), 923–931.

Rauschenberger, M., Rello, L., & Baeza-Yates, R. (2025). Screening dyslexia using visual auditory computer games and machine learning. *IEEE Access*, 13.

Rello, L., & Ballesteros, M. (2019). Predicting risk of dyslexia with an online gamified test. *PLoS ONE*, 14(11), e0224879.

Richlan, F., Kronbichler, M., & Wimmer, H. (2013). Structural abnormalities in the dyslexic brain: A meta-analysis of voxel-based morphometry studies. *Human Brain Mapping*, 34(11), 3055–3065.

Robaa, M., Balat, M., Awaad, R., Omar, E., & Aly, S. A. (2024). Explainable AI in handwriting detection for dyslexia using transfer learning. *arXiv*.
<https://arxiv.org/abs/2410.19821>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Shaywitz, S. E., Morris, R., & Shaywitz, B. A. (2008). The education of dyslexic children from childhood to young adulthood. *Annual Review of Psychology*, 59, 451–475.

Shifrer, D., Muller, C., & Callahan, R. (2011). Disproportionality and learning disabilities: Parsing apart race, socioeconomic status, and language. *Journal of Learning Disabilities*, 44(3), 246–257.

Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., ... Ashrafian, H. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, 31(10), 3283–3289.

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer.

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 106, 359–380.

Torppa, M., Georgiou, G., Salmi, P., Eklund, K., & Lyytinen, H. (2012). Examining the double-deficit hypothesis in an orthographically consistent language. *Scientific Studies of Reading*, 16(4), 287–315.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., ... McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5), 924–933.

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology*, 91(3), 415–438.

Zugarramurdi, C., Fernández, L., Lallier, M., Carreiras, M., & Valle-Lisboa, J. C. (2022). Lexiland: A tablet-based universal screener for reading difficulties in the school context. *Journal of Educational Computing Research*, 60(7), 1688–1715.