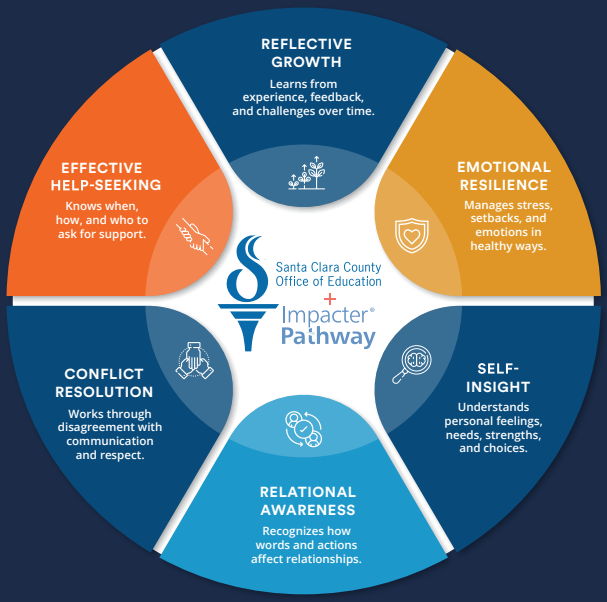
From Self-Report to Student Voice

A Guide to Performance-Based Behavioral Health Screening

How IMPACTER measures behavioral health through authentic student voice —the research foundation, the rubric-aligned scoring system, and the evidence behind it.



For district behavioral health leaders, school psychologists, and county MTSS teams.

Impacter® Pathway

Version 2.1

1. An Authentic-Voice Approach to Behavioral Health Screening

IMPACTER measures human skills through authentic student voice. Students respond to open-ended prompts in their own words. A rubric-aligned scoring engine — trained directly on expert human raters working with validated rubrics — analyzes the language of those responses. The result is a developmental, rubric-aligned score that routes each student to the right level of behavioral health support within a Multi-Tiered Systems of Support (MTSS) framework.

This guide describes how that works.

It starts with the current screener landscape and the gaps that performance-based methods are well-suited to fill. It then walks through the research foundation underneath authentic-voice measurement, the operational practices — rubrics, expert human raters, machine learning, ongoing calibration — that make this measurement reliable at scale, and the three independent layers of evidence that support its use as a behavioral health screener

The audience is district behavioral health leaders, school psychologists, county MTSS teams, and anyone responsible for evaluating screening tools on their measurement merits.

2. The Current Screener Landscape

Schools use behavioral health screeners to identify students who may benefit from additional support. The most commonly used K–12 instruments — PHQ-9 for depression, GAD-7 for anxiety, the BASC-3 Behavioral and Emotional Screening System for broad behavioral and emotional concerns, the Columbia–Suicide Severity Rating Scale for suicidal ideation, and survey-based tools like Panorama and Aperture/DESSA — share a common methodology: a fixed set of items, presented to every student, rated on a Likert scale.

This methodology is well-suited to many use cases. Items are short. Administration is fast. Scoring is straightforward. And for several of these instruments, particularly the PHQ-9 and GAD-7, decades of validation research link specific item responses to clinical outcomes (Kroenke et al., 2001; Spitzer et al., 2006; Levis et al., 2019; Posner et al., 2011).

WHERE THE GAPS EMERGE.

Likert-scale self-report has well-documented limitations in K–12 settings — limitations that have been established across multiple research literatures and are summarized briefly here:

- **Social desirability bias.** Students tend to select answers they think will be perceived favorably, particularly on items with strong social meaning (Paulhus & Vazire, 2007).
- **Response set bias.** Adolescents show patterned response behavior — straight-lining, mid-point preference, ceiling responses — that compresses variance and obscures real differences (Krosnick, 1991).
- **Ceiling effects on strengths-based items.** Adolescent populations frequently cluster near the top of the scale on positive items, making the instrument insensitive to the variance it was designed to detect.
- **Vocabulary asymmetry.** Self-report items assume the respondent already has language for the construct being rated — an assumption that fails systematically for younger students, English Learners, and students with limited prior exposure to social-emotional vocabulary.
- **Categorical, not developmental.** Likert-based screeners typically produce a categorical signal (at risk / not at risk). They were designed to identify symptoms, not measure growth. For MTSS-aligned systems where the goal is matching support to evidence of need and tracking change over time, a developmental signal is more useful than a binary one.

These are not flaws to be apologized for. They are properties of a methodology that does some things very well and was never designed to do others. The field benefits from instruments that complement Likert-based screeners rather than replace them.

The Screener Landscape

Instrument	Modality	Elicitation Method	Validity Approach	MTSS-Ready	Language Content Captured	Behavioral Health Integration
IMPACTER	Performance-Based	Authentic Voice	Rubric Validation	✓	✓	✓
Panorama	Likert	Uniform Prompt	ItemValidation	✓	✗	✓
Aperture/DESSA	Likert	Uniform Prompt	ItemValidation	✓	✗	✓
BASC-3 BESS	Likert	Uniform Prompt	ItemValidation	✓	✗	✓
PHQ-9	Likert	Uniform Prompt	ItemValidation	✗	✗	✓

✓ Yes | ✗ No

The figure shows where IMPACTER sits in the landscape: performance-based modality, authentic student voice as the elicitation method, rubric validation as the validity approach, and language content captured directly — alongside MTSS readiness and behavioral health integration matched to the established Likert-based instruments.

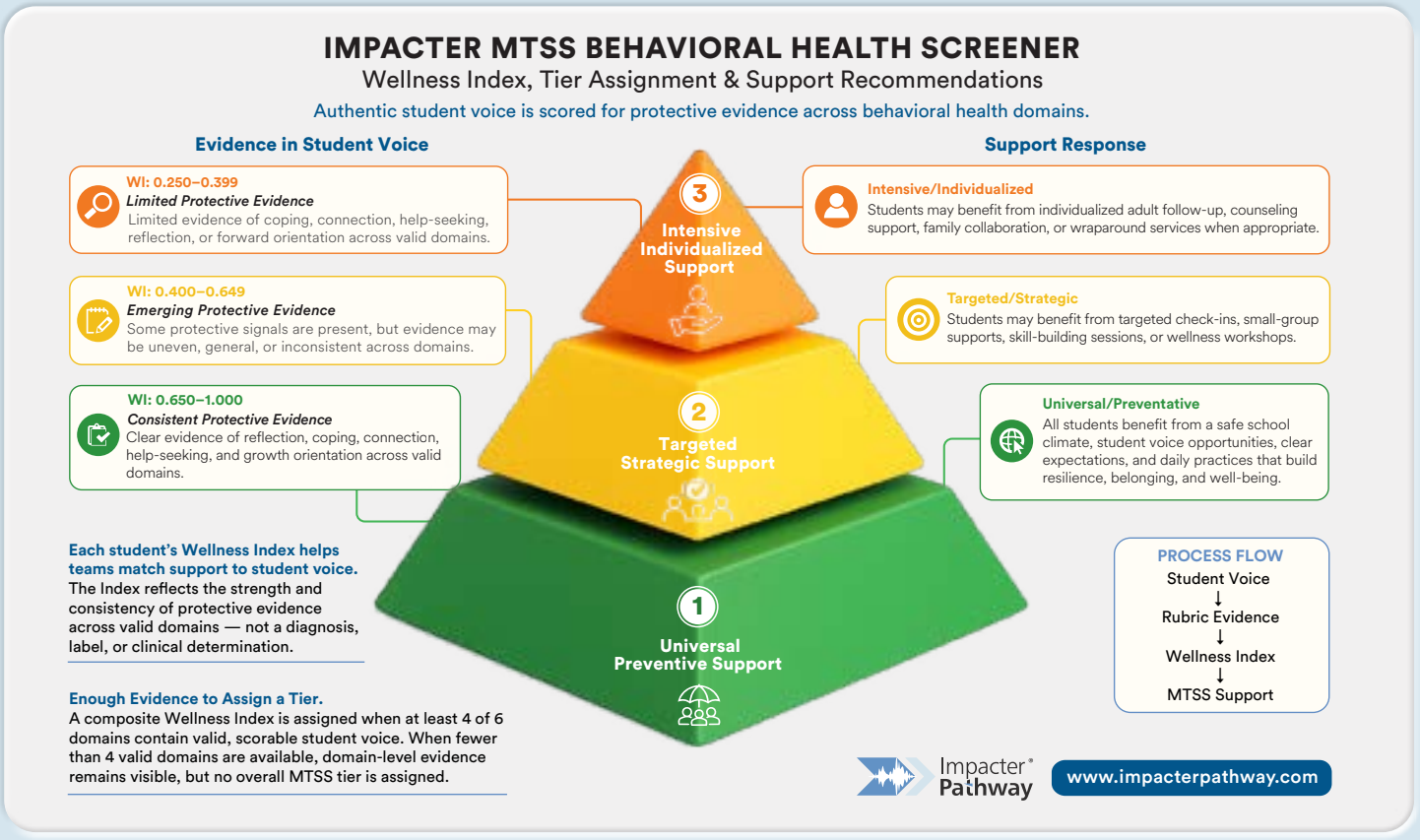
3. How IMPACTER Closes the Gap

IMPACTER addresses each of the gaps above through a single methodological move: rather than asking students to rate themselves on traits, IMPACTER asks students to respond to prompts in their own words — and measures the language of those responses against validated rubrics.

This changes what is captured:

- **Authentic voice replaces self-rating.** Students describe experiences in their own language. The measurement keys off what the response actually demonstrates, not what the student claims about themselves.
- **Performance-based replaces categorical.** Each response is scored on a developmental rubric (Foundational → Developing → Competent → Advanced). The signal carries information about where a student is in their development of a competency, not just whether they cross a threshold.
- **Direct construct content replaces inference.** The linguistic features of an authentic response — specificity, perspective-taking, coping strategies, help-seeking language — are observable evidence of the underlying construct. There is no need to infer that a student has emotional regulation skills from a self-rating; the response itself either does or does not demonstrate them.
- **Growth signal replaces snapshot.** Because the rubric produces ordinal developmental scores, a student's measurement on the same competency this year vs. last year is directly comparable across the full range — not just at the threshold.

The output is the Wellness Index — a composite of rubric-aligned scores across six behavioral health domains, routed to one of three MTSS tiers: Tier 1 (Universal/Preventive), Tier 2 (Targeted/Strategic), or Tier 3 (Intensive/Individualized). The Wellness Index reflects the strength and consistency of protective evidence across valid domains. It is not a diagnosis, a label, or a clinical determination — it is rubric-aligned evidence used to match level of support to level of need.



The Wellness Index supports professional judgment; it does not replace it. Tier 2 and Tier 3 decisions are made by adult teams using IMPACTER's evidence alongside teacher observation, attendance, behavioral patterns, and family communication — the standard of practice for any behavioral health screener.

4. How IMPACTER Enters the Screener Space

A behavioral health screener — across medical, behavioral health, and educational contexts — is a tool that meets five established criteria. It is brief. It is broadly administrable. It identifies rather than diagnoses. It is reasonably sensitive and specific. And it produces reliable, comparable results. The U.S. Preventive Services Task Force and the AERA/APA/NCME Standards for Educational and Psychological Testing converge on this functional definition.

These criteria describe what a screener has to do — identify, briefly, universally, reliably. They do not specify how a screener has to elicit a response. Different categories of validated screeners satisfy them through different mechanisms.

IMPACTER meets each criterion:

Figure 4

Screener Criteria Checklist

Section 2

SCREENER CRITERION	HOW IMPACTER IMPLEMENTS IT	
Brief	Uses short, focused prompts (3–5 minutes) that respect instructional time while yielding meaningful signals.	✓
Broadly Administrable	Designed for universal administration across grades, settings, and student populations at scale.	✓
Identifies, Does Not Diagnose	Flags students who may benefit from additional support; it does not diagnose conditions.	✓
Reasonably Sensitive and Specific	Scores are validated against human raters and show strong agreement and meaningful differentiation.	✓
Standardized for Reliability	Standardizes the measurement layer through fixed rubrics, scoring logic, balanced prompt sets, and model versioning.	✓

IMPACTER®

Source: U.S. Preventive Services Task Force screening criteria; AERA/APA/NCME Standards for Educational and Psychological Testing.

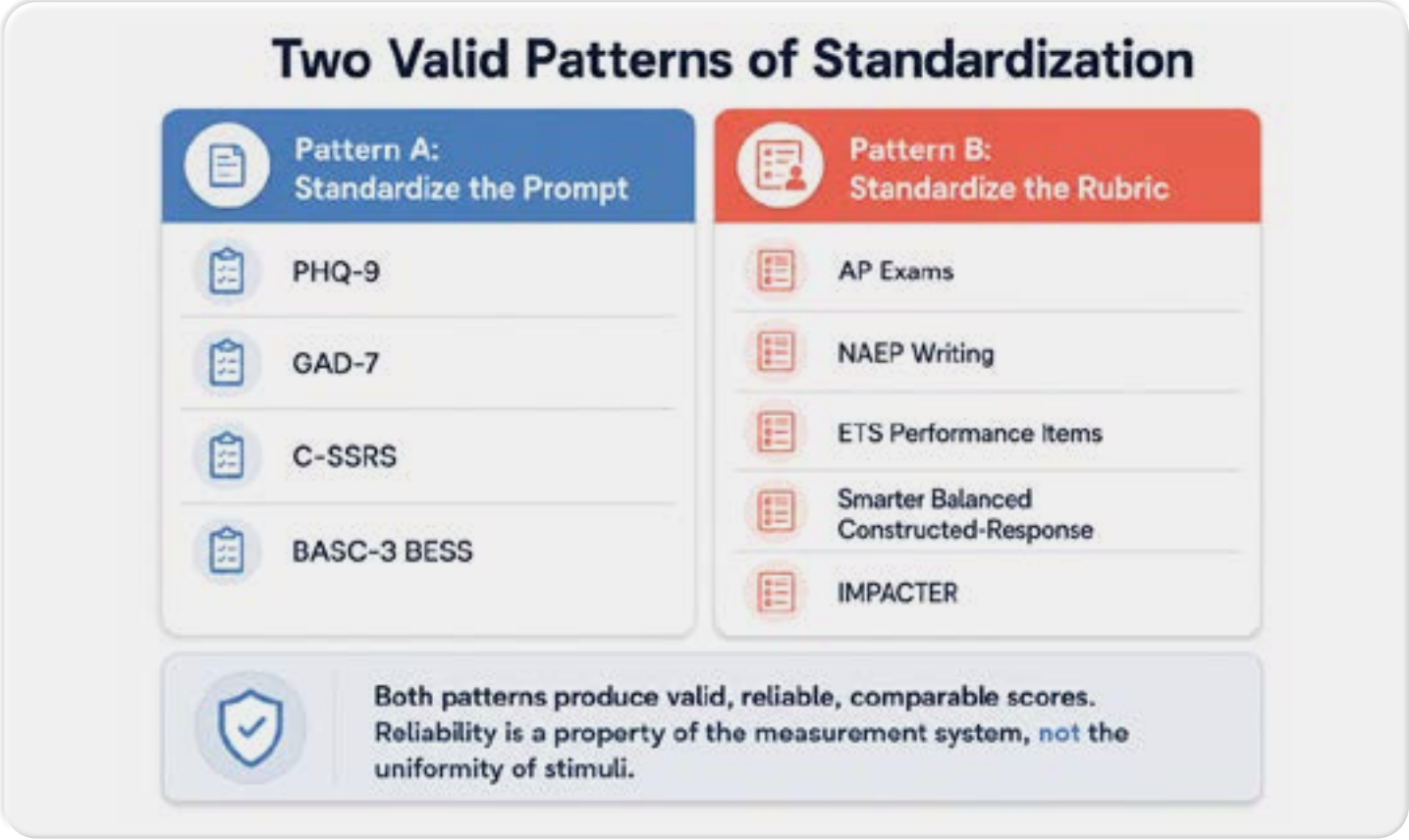
IMPACTER also matches the operational discipline expected of a screener: documented thresholds for promotion and drift detection, scoped intended use, explicit out-of-scope conditions (high-stakes summative academic scoring, individual diagnostic or clinical decisions without human review, and any use outside rubric-anchored prompts), and a versioned Model Card that documents performance per release. A screener identifies; it does not diagnose; IMPACTER is documented in scope and out of scope accordingly.

5. How IMPACTER Standardizes Measurement

Validated assessments achieve standardization through one of two patterns.

- **Pattern A: Standardize the prompt.** PHQ-9, GAD-7, C-SSRS, BASC-3 BESS, and similar instruments present every respondent with the same items, scored against a fixed key. Reliability comes from uniformity of stimuli. Validity is established by demonstrating that responses to those specific items, asked in that specific way, correlate with the target construct or outcome (Posner et al., 2011; Kroenke et al., 2001).
- **Pattern B: Standardize the rubric.** AP Exams, NAEP Writing, ETS Performance Items, Smarter Balanced constructed-response components, and IMPACTER present different students with different prompts — and score every response against the same rubric, applied by trained human raters or rubric-aligned scoring engines validated against trained human raters. Reliability comes from uniformity of measurement. Validity is established by demonstrating that rubric-aligned scores reflect the underlying construct and correlate with intended outcomes (Messick, 1989; Kane, 2013).

Both patterns produce valid, reliable, comparable scores. AP US History does not become invalid because the DBQ topic changes year to year — the validity lives in the scoring system.



IMPACTER operates in Pattern B. Different students may receive different prompts. Every response is scored against the same rubric, by the same rubric-aligned scoring engine, with deterministic inference (IMPACTER Model Card v6.0). The standardization lives in the measurement, which is exactly where it needs to live for a Pattern B instrument.

Pattern B instruments do not treat prompts as irrelevant — they treat prompts as a studied facet of measurement, monitored through generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001). IMPACTER's prompt families are engineered for equivalence within rubric-anchored task types: aligned to the same competency targets, calibrated against the same rubric anchors, and audited for consistency in construct content. Prompt equivalence is treated as ongoing measurement discipline, not a one-time validation event.

6. The Research Foundation

Section 5 establishes that standardizing the rubric is a valid path to comparable scores. Section 6 establishes what is underneath the rubric: a body of psycholinguistic research showing that specific linguistic features are observable markers of specific cognitive and emotional capacities.

IMPACTER did not invent this link. The research connecting language to human capacities long predates the company. What IMPACTER does is operationalize that research at scale, against rubrics co-developed with Harvard's Making Caring Common Project.

EMOTIONAL GRANULARITY.

Lisa Fveldman Barrett's research program (Barrett, 2017; Barrett, Gross, Christensen, & Benvenuto, 2001) demonstrates that individuals who differentiate emotional states with greater linguistic precision regulate those states more effectively, experience fewer mental health symptoms, and respond more adaptively to stress. The capacity to label and articulate emotions in fine-grained terms is not merely a description of emotion regulation — it is one of its operative mechanisms. Mark Brackett's work at the Yale Center for Emotional Intelligence (Brackett, 2019) extends this finding into school settings: students who can name what they are feeling, identify why, and articulate strategies to address it demonstrate stronger self-regulation, healthier peer relationships, and better academic outcomes.

THEORY OF MIND.

Four decades of research (Premack & Woodruff, 1978; Wellman, Cross, & Watson, 2001) establish that specific linguistic features — mental-state verbs ("she thought," "he believed"), pronoun-perspective shifts ("from her view," "I could see why he..."), embedded clauses describing internal states — are observable markers of perspective-taking capacity. The Wellman, Cross, & Watson meta-analysis of 178 false-belief studies confirms that the capacity to represent other minds and the linguistic markers of that capacity develop in parallel, and are detectable from early childhood through adolescence.

GROWTH MINDSET.

Dweck's research program (Dweck, 2006; Yeager & Dweck, 2012; Yeager et al., 2019) identifies effort-attribution statements, incremental-change verbs, and contrastive structures ("I used to ... but now I can ...") as observable markers of an incremental theory of intelligence.

PURPOSE AND MOTIVATION.

Self-Determination Theory (Ryan & Deci, 2000) identifies goal-framing language, value statements, and abstract mission phrasing as markers of intrinsic motivation and motivational coherence.

GRIT.

Duckworth's research (Duckworth, 2016; Duckworth, Peterson, Matthews, & Kelly, 2007) identifies temporal sequencing of effort, adversity-persistence narratives, and endurance verbs as markers of sustained goal pursuit.

PROSOCIAL DEVELOPMENT.

Research on prosocial reasoning and reciprocity (Eisenberg, Fabes, & Spinrad, 2006) identifies benefit-acknowledgment language, joint-agency phrasing, and reciprocity statements as markers of relational awareness.

In each case, the linguistic feature is not a metaphor for the capacity. It is an observable surface marker, established in primary research. The rubric IMPACTER scores against is a careful operational mapping from those research-established markers to scorable rubric levels.

7. How IMPACTER Applies the Research

Translating decades of psycholinguistic research into reliable measurement at the scale of a district behavioral health screener requires four operational practices: rubric-aligned training with expert human raters, a rubric-anchored machine learning pipeline, an ongoing inter-annotator agreement system, and a versioned model registry with deterministic inference.

7.1 RUBRIC-ALIGNED TRAINING WITH EXPERT HUMAN RATERS.

The training signal for IMPACTER's scoring engine comes from expert human raters — educators, counselors, and assessment specialists working with dual-axis rubrics that define observable developmental levels for each competency. These rubrics are co-developed with Harvard's Making Caring Common Project and cross-walked to the CASEL competency taxonomy, so districts working in either framework get coverage. Expert human ratings on real student responses are the labels the scoring engine learns from.

sd coe

BEHAVIORAL HEALTH CROSSWALK

Prompt Title	Behavioral Health Focus (Student)	IMPACTER Attribute(s)	Reflection Themes	Observable Indicators	MTSS Domain (System/Adult)	CA Dashboard Outcome
1. That Wasn't Really Me	*Self-Insight*		Understanding self, recognizing choices, describing moments of growth, taking responsibility, connecting actions to values	Student shares a story of acting out of character, explains what they learned, describes how they want to act in the future	Whole Child Domain	Attendance, Student Behavior
2. The Hard Days	*Emotional Resilience*		Naming emotions, using self-care or coping strategies, perseverance through difficulty, learning what helps them cope	Student describes a difficult moment, identifies their feelings, discusses strategies they used, and shows perseverance	Integrated Supports Domain	Attendance, Academic Achievement
3. Something Changed	*Relational Awareness*		Noticing changes in relationships, understanding impact, expressing empathy, valuing connection, learning from shifts	Student shares about a relationship change, demonstrates empathy, describes how they adapted or grew from the experience	Family and Community Engagement Domain	Student Engagement/ Climate, English Proficiency
4. Fixing What Broke	*Conflict Resolution*		Taking responsibility, practicing repair, showing persistence, understanding consequences, striving to make things better	Student describes a repair attempt, articulates steps taken, shows understanding of outcomes, reflects on persistence	Inclusive Policy Structure & Practice Domain	Student Behavior, Graduation Rate
5. Stronger Together	*Effective Help-Seeking*		Knowing when and how to ask for help, using available supports, recognizing teamwork, being open to guidance	Student talks about asking for help, names who they turned to, describes positive outcomes from support	Administrative Leadership Domain	Academic Achievement, Attendance
6. Not the Same Person	*Reflective Growth*		Describing personal growth, identifying change, setting goals, reflecting on progress, believing in improvement	Student gives examples of change, identifies what contributed to growth, expresses confidence in continued growth	Whole Child Domain	Graduation Rate, Academic Achievement

Impacter Pathway

sd coe

BEHAVIORAL HEALTH RUBRIC

Impacter Pathway

The crosswalk and rubric together do operational work: they specify exactly what evidence in a student response indicates each developmental level, for each competency, in language that human raters and the scoring engine can apply consistently.

7.2 A RUBRIC-ANCHORED MACHINE LEARNING PIPELINE.

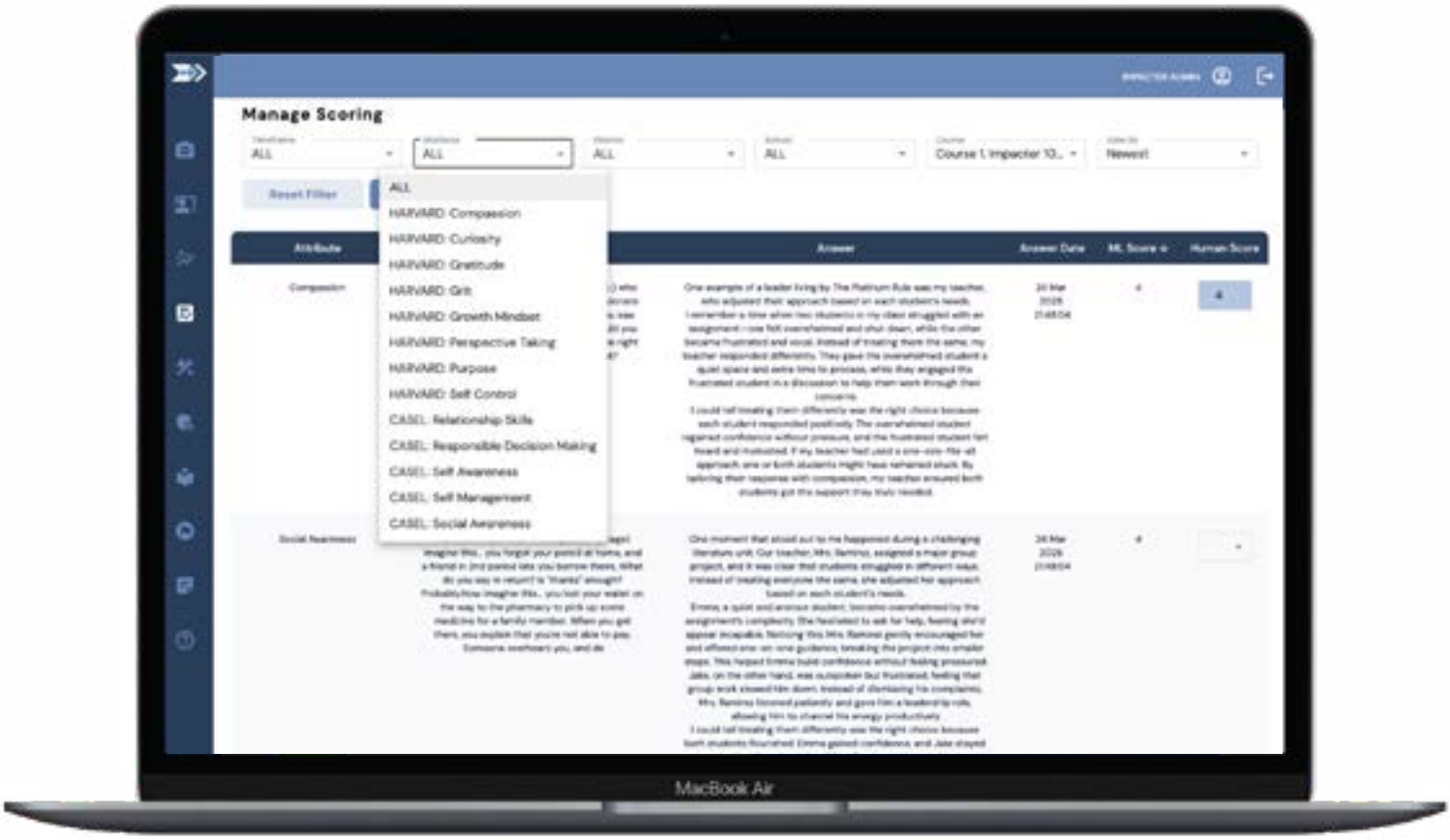
IMPACTER's scoring engine is built on a fine-tuned DeBERTa-v3-base transformer with a CORAL ordinal regression head, trained directly on expert-rated student responses. The architecture matters. Standard classification treats all errors equally, but CORAL enforces rank structure: the model learns that a Level 2 response is closer to Level 1 than to Level 0. The loss function penalizes predictions in proportion to their distance from the true rating. For developmental constructs like grit or perspective-taking, this structurally prevents the model from confusing "Emerging" with "Exemplary".

This is a discriminative pipeline, not a generative one. The model encodes a response and produces an ordinal score directly. There is no text generation, no token-by-token sampling, no stochastic element at inference. Identical inputs produce identical scores across every administration.

This distinction matters in the broader scoring methodology context. Research published in the Journal of Educational Data Mining (Ober et al., 2026) examined generic large language models scoring student dialogues for non-cognitive constructs. The finding: prompted general-purpose models showed high model-to-model agreement but low model-to-human agreement — internally consistent, but systematically diverging from expert human judgment. The diagnosis offered is that prompted generative models perform in-context pattern matching on surface linguistic cues without grounded training signal from human raters scoring against an actual rubric. IMPACTER's discriminative architecture, trained directly on rubric-aligned human ratings, is the methodological alternative the research points toward. The model is not asked to invent an evaluation framework on the fly; it learns the framework from the human ratings that define it.

7.3 EXPERT-VALIDATED RATERS AT SCALE: THE INTER-ANNOTATOR AGREEMENT SYSTEM.

Human-in-the-loop calibration does not end at training. IMPACTER's Admin Platform includes an Inter-Annotator Agreement (IAA) interface where trained human raters review machine-learning-generated scores alongside their own, flag disagreements, and feed corrections back into the training cycle. The interface tracks score comparisons at the individual response level, filterable by competency, district, school, and timeframe.



This is the operational equivalent of ongoing inter-rater reliability monitoring in traditional rubric-aligned assessment — adapted for a system in which one of the raters is a model whose performance is continuously calibrated against the other. Beyond reliability, the IAA system surfaces a useful diagnostic signal: where machine-human disagreement concentrates around particular constructs, prompt families, or response patterns, those concentrations point back to rubric refinement, model retraining, or construct redefinition.

7.4 VERSIONING, AUDIT, AND DETERMINISTIC PRODUCTION.

Once a model version meets agreement and fairness thresholds (QWK ≥ 0.70, adjacent agreement ≥ 94%, |SMD| < 0.15, |SMD|s < 0.10), it is registered in MLflow and promoted to Production. Production models are frozen in the registry. Identical responses receive identical scores across administrations. Intended use and out-of-scope conditions are explicitly published in the IMPACTER Model Card.

MODEL CARD

Version 6.0 | CORAL Ordinal Regression Models

Impacter Pathway

Overview & Scope

This model suite scores K-12 student open-ended responses on eight human-skills competencies using a fine-tuned DeBERTa-v3-base architecture. Models generate rubric-aligned ordinal scores calibrated to human expert scoring of student reasoning, reflection, and narrative evidence.

Models are trained on annotated student responses from multiple school district implementations and evaluated using standard educational measurement agreement statistics (QWK, adjacent agreement, SMD).

Intended Use

- Scoring student reflections for Portrait of a Graduate / soft-skills competencies
- Behavioral health and social-emotional reflection scoring
- Workforce and career readiness / CTE assessments
- Short-form constructed-response tasks

Out of Scope

- High-stakes summative academic content scoring
- Individual diagnostic or clinical decisions without human review
- Any use outside of rubric-anchored prompts and approved task types

Training & Evaluation Data

The model is trained on district-provided student responses of written and transcribed speech paired with hand-scored human evaluations on a 0-4 scale. These examples come from authorized educational partners and represent authentic answers to open-ended prompts. Scoring is performed on newly collected student responses using held-out validation sets.

Model Configuration

Models are trained using CORAL ordinal regression and evaluated on held-out data. New versions are registered in MLflow and promoted to Production only if QWK improves relative to the current Production model.

Once promoted, Production models are frozen in the registry for deterministic scoring and full auditability. All inference uses frozen Production models with deterministic tokenization and scoring, ensuring identical responses receive identical scores across administrations.

Model releases incorporate SMD and subgroup fairness checks aligned to educational standards (QWK ≥ 0.70, degradation ≤ 0.10, |SMD| < 0.15, $|SMD|_s$ < 0.10).

Model Configuration

- **Architecture:** Microsoft/ DeBERTa-v3-base (184M parameters)
- **Version:** IMPACTER Model Suite v6.0
- **Loss Function:** CORAL Ordinal Regression
- **Platform:** Databricks with MLflow Model Registry

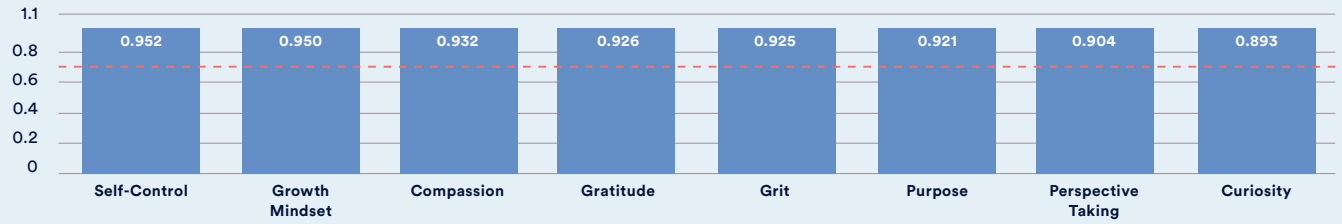
Summary Statistics

- **QWK (Quadratic Weighted Kappa):** Measures human-ML scoring agreement with higher weight on larger disagreements. Range: -1 to 1, where 1 = perfect agreement.
- **Accuracy:** Percentage of ML scores that exactly match human scores. Measures precise scoring alignment.
- **Adjacent Accuracy:** Percentage of ML scores within ±1 point of human scores. Typical human-human rater agreement benchmark.

AVG QWK	AVG ACCURACY	AVG ADJACENT ACC
0.918	77.3%	98.0%

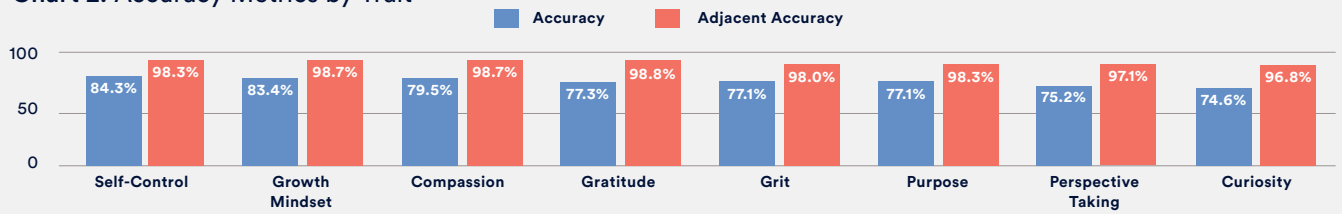
Data Visualization

Chart 1: QWK Performance by Trait



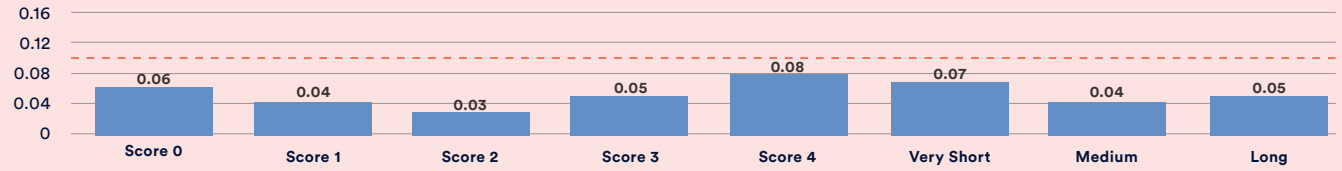
Red line indicates industry-standard minimum ML scoring threshold (QWK ≥ 0.70). All traits exceed this threshold.

Chart 2: Accuracy Metrics by Trait



Exact accuracy ranges from 75–84% while adjacent agreement (±1) ranges from 97–99% across traits, consistent with human-rater monitoring expectations.

Chart 3: Preliminary Fairness Check (Structural SMD)



Preliminary distributional SMD checks (by score/length bands) show no systematic drift.
Full subgroup SMD analysis (EL status, race/ethnicity) will be computed using partner-specific human-scored validation sets.

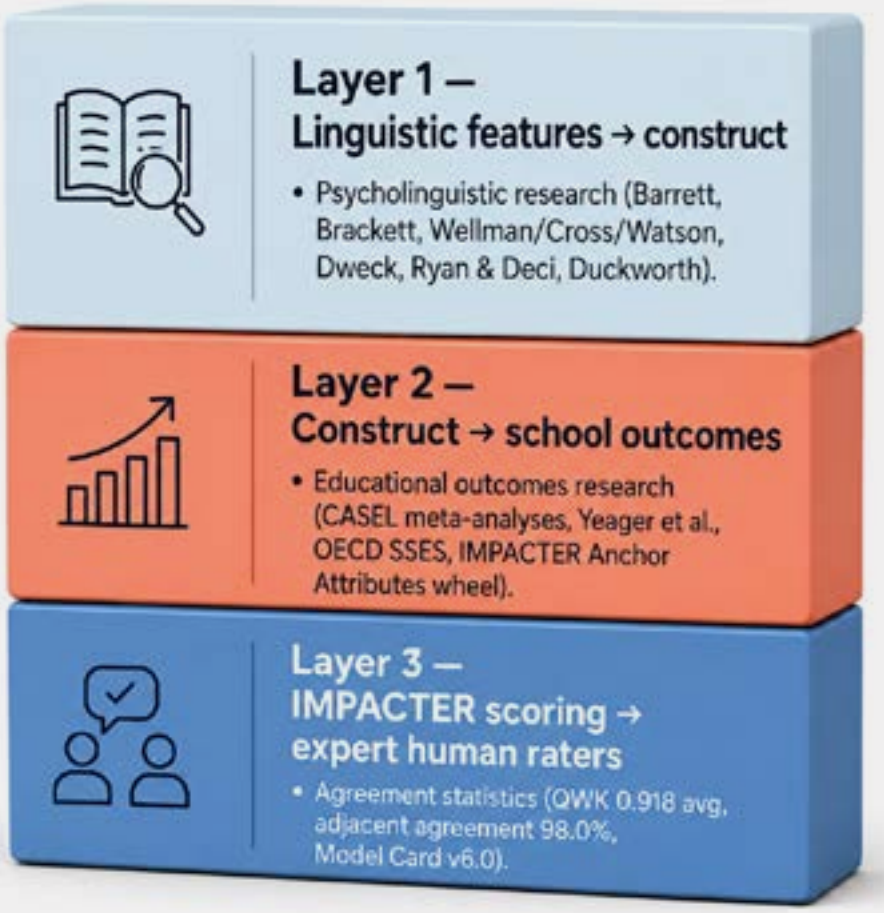
IMPACTER Pathway Assessment | DeBERTa CORAL Model Training Report v6.0
All models meet industry-standard thresholds for automated scoring (QWK ≥ 0.70, Adjacent Agreement ≥ 94%)
Models promoted to Production in MLflow with frozen inference for deterministic scoring

This versioning and audit infrastructure is the operational equivalent of the publication and revision history of a validated screening scale. Where the PHQ-9 has documented item history and validation studies (Kroenke et al., 2001; Levis et al., 2019), IMPACTER has a versioned model registry, frozen inference, a published Model Card with reproducible performance statistics, and a continuously operating human-in-the-loop calibration system.

8. Three Layers of Validity Evidence

Validity for any assessment instrument is established by triangulating across multiple sources of evidence (Messick, 1989; AERA/APA/NCME, 2014). For IMPACTER, three independent layers of evidence converge on the same conclusion.

Figure 2: Three Layers of Validity Evidence



Three independent sources of evidence. Each established without reference to the others. All converge on the same conclusion.

LAYER 1: LINGUISTIC FEATURES → CONSTRUCT.

The psycholinguistic research summarized in Section 6 establishes the link between specific linguistic features and specific human capacities. Across emotional granularity, theory of mind, growth mindset, self-determination theory, grit research, and prosocial development, the links are foundational findings — not invented for IMPACTER, but operationalized by IMPACTER's rubrics.

Third-party validation of the architectural approach has been published independently. Ober and colleagues at ETS Research (Ober et al., 2026) demonstrate that generic large language models, applied without rubric-aligned discriminative training, fail to produce reliable scoring of constructed-response items. Rubric-aligned, discriminatively trained models — the IMPACTER architectural pattern — meet the agreement thresholds required for valid automated scoring. This is independent validation of the methodology by a research group with no commercial relationship to IMPACTER.

LAYER 2: CONSTRUCT → SCHOOL OUTCOMES.

The second layer is the documented relationship between the human-skill constructs IMPACTER measures and measurable school outcomes — relationships established in the broader research literature, not invented for IMPACTER. The IMPACTER Anchor Attributes and Measurable Outcomes mapping (IMPACTER Pathway, 2024) traces these links:

- **Curiosity →** academic achievement (higher GPAs, decreased D/F rates)
- **Perspective-taking →** English proficiency and language reclassification rates
- **Purpose →** attendance and tardy/truancy rates
- **Self-control →** reduced suspension and expulsion rates
- **Grit →** graduation rates and post-secondary attainment
- **Growth mindset →** improved performance on standardized assessments
- **Compassion →** school climate survey outcomes and participation in school activities
- **Gratitude →** enrollment in advanced coursework

These links draw on Yeager and colleagues at Stanford and the University of Texas (Yeager et al., 2019), the CASEL meta-analyses (Durlak et al., 2011; Mahoney et al., 2018), the OECD Survey on Social and Emotional Skills (OECD, 2021), and decades of attendance and school climate research. The constructs IMPACTER measures are well-established predictors of student trajectories.

LAYER 3: IMPACTER SCORING → EXPERT HUMAN RATERS.

The third layer is documented agreement between IMPACTER's automated scoring and expert human raters on the same student responses. The current production model suite (v6.0) achieves an average Quadratic Weighted Kappa (QWK) of 0.918 across the eight measured traits, with per-trait QWK ranging from 0.893 to 0.952. Average adjacent agreement (scores within ±1 of human raters) is 98.0%, with per-trait adjacent agreement ranging from 96.8% to 98.8%.

These figures exceed industry-standard thresholds for automated scoring of constructed-response items (Williamson, Xi, & Breyer, 2012; ETS, 2018): QWK ≥ 0.70 and adjacent agreement ≥ 94%. Adjacent agreement is the more meaningful comparison metric for ordinal scoring, since trained human raters themselves typically achieve adjacent rather than exact agreement on rubric-aligned tasks.

Per-construct agreement statistics also serve as a diagnostic for construct clarity. Self-Control (QWK 0.952) has the tightest rubric anchors and lowest scoring divergence; Curiosity (QWK 0.893) is multidimensional — bundling information-seeking, ambiguity tolerance, intrinsic motivation, and cross-domain transfer — and produces relatively more divergence, identifying it as a construct where rubric tightening remains an ongoing priority.

The three layers do not depend on one another to be true. Each is established by independent evidence — psycholinguistic research, educational outcomes research, and direct agreement studies. They converge on the same conclusion: IMPACTER measures what it claims to measure.

FAIRNESS AND DRIFT MONITORING.

IMPACTER operates under documented fairness thresholds aligned to educational measurement standards. Structural fairness checks across score distributions and response-length distributions are completed and meet published thresholds, with no systematic drift detected. Demographic subgroup fairness analyses (English Learner status, race/ethnicity) are computed against partner-specific human-scored validation sets at each implementation — making fairness an ongoing operational discipline tied to local data, rather than a one-time global validation event.

9. A Note on the Limits of Self-Report

Section 2 introduces the gaps in Likert-scale self-report. This section returns to those gaps in more depth — not to argue against self-report screeners, which serve a real purpose, but to explain why authentic-voice methodology captures signal that Likert is structurally unable to capture.

Social desirability bias is well-documented in the survey methodology literature (Paulhus & Vazire, 2007). When students are asked to rate themselves on "I am a hard worker" or "I can handle my emotions," they tend to select responses they believe will be perceived favorably — particularly when items are face-valid for traits with strong social meaning. This is amplified in school settings, where students often anticipate that low responses may have consequences.

Response set bias compounds the problem. Adolescents in particular show patterned response behavior — straight-lining, ceiling responses, mid-point preference — that compresses variance and obscures real differences between students (Krosnick, 1991). The instrument can produce clean numerical data while still failing to detect the variance it was designed to capture.

Ceiling effects are a structural property of strengths-based items in adolescent populations. On positive items, students frequently cluster near the top of the scale, making the instrument insensitive to differences that matter for screening and growth measurement.

And there is a vocabulary asymmetry that the field has documented for decades. Self-report items assume the respondent already has language for the construct being rated. For younger students, English Learners, and students with limited prior exposure to social-emotional vocabulary, that assumption fails systematically. Students who do not yet have a word for "perseverance" cannot reliably rate themselves on it.

IMPACTER's authentic-voice methodology captures signal that Likert self-report is structurally unable to capture. A student who has never been asked to articulate what perseverance looks like in their ownlife — and whose response, in their own words, surfaces vivid evidence of it — is providing measurement information that no rating scale can produce. The instrument meets students where their language actually lives, rather than asking them to translate their experience into someone else's scale.

This is not an argument against self-report screeners. It is an argument that the field benefits from complementary methodologies, and that performance-based measurement closes a measurement gap that has been documented for decades.

10. Ongoing Validation

Validity is an argument built from accumulating evidence, not a settled fact (Messick, 1989; Kane, 2013). The evidence supporting IMPACTER's use as a performance-based screener is substantial and continues to develop. Several workstreams are actively underway, each adding evidence to the validity argument:

- **Generalizability theory studies.** Partitioning measurement variance across student, trait, prompt family, modality, and rater facets, with reportable variance components published for each major task type.
- **Classification performance evaluation.** Sensitivity, specificity, positive predictive value, and negative predictive value at published cut points, evaluated against partner-defined outcomes, with referral burden monitored against district capacity.
- **External convergence and divergence studies.** Relations between IMPACTER scores and established behavioral health indicators, academic performance measures, and theoretically unrelated constructs — including demonstrating that IMPACTER scores are not reducible to general writing proficiency.
- **Response-process evidence.** Student cognitive interviews and rater cognitive walkthroughs to confirm that prompts elicit the intended construct content and that rubric-aligned human scoring keys off intended evidence rather than surface features.
- **Subgroup fairness audits.** Demographic subgroup analyses (English Learner status, race/ethnicity, gender) at each partner implementation, with documented thresholds and drift monitoring.

- **Longitudinal predictive validity.** Relations between IMPACTER scores and downstream school outcomes — attendance, behavioral incidents, graduation, post-secondary trajectories — in partner cohorts.
- **Inter-annotator agreement at scale.** Continued expansion of the human-rater calibration system across competencies, prompt families, and partner contexts, with published agreement statistics in each annual Model Card release.

Several of these workstreams are being developed in collaboration with academic and county partners, including the Behavioral Health Student Services Act (BHSSA) evaluation led by the Santa Clara County Office of Education, with Stanford University and Applied Survey Research as research partners. IMPACTER's validation operates as a publicly accountable, partnership-based research program rather than a proprietary internal exercise.

As each workstream reaches publication, results are incorporated into the IMPACTER Model Card and into updated versions of this guide. The current document reflects the state of the evidence as of the publication date; the underlying research program is continuous.

11. Conclusion

This guide describes how IMPACTER measures behavioral health through authentic student voice — the research foundation underneath the method, the rubric-aligned scoring system that operationalizes it, the inter-annotator agreement system that calibrates it continuously, and the three independent layers of evidence that support its use as a behavioral health screener.

The instrument fills a documented gap in the K–12 screener landscape: a developmental, performance-based signal that captures what students actually say, in their own words, against rubrics that map to the human capacities most predictive of school and life trajectories. It does this within the operational discipline expected of any serious measurement instrument: versioned scoring, frozen production inference, published thresholds, scoped intended use, and ongoing partner-specific validation.

We welcome direct dialogue with district psychometricians, county behavioral health teams, and academic researchers. Validation is an ongoing practice, and IMPACTER's commitment to versioned auditing, published thresholds, partner-specific fairness analyses, and continuous human-in-the-loop calibration reflects that posture.

References

SCREENING METHODOLOGY AND VALIDITY

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Brennan, R. L. (2001). Generalizability theory. Springer-Verlag. https://doi.org/10.1007/978-1-4757-3456-0
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Wiley.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). American Council on Education.

U.S. Preventive Services Task Force. (2021). Procedure manual. https://www.uspreventiveservicestaskforce.org/uspstf/about-uspstf/methods-and-processes/procedure-manual

VALIDATED SCREENERS CITED

Kamphaus, R. W., & Reynolds, C. R. (2015). BASC-3 Behavioral and Emotional Screening System (BESS). Pearson. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/basc-3/basc3-bess-overview.pdf

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression. BMJ, 365, l1476. https://doi.org/10.1136/bmj.l1476

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., et al. (2011). The Columbia–Suicide Severity Rating Scale. American Journal of Psychiatry, 168(12), 1266–1277. https://doi.org/10.1176/appi.ajp.2011.10111704

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

PSYCHOLINGUISTIC FOUNDATIONS

Barrett, L. F. (2017). How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt. https://www.hmhbooks.com/shop/books/How-Emotions-Are-Made/9781328915436

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it. Cognition and Emotion, 15(6), 713–724. https://doi.org/10.1080/02699930143000239

Brackett, M. A. (2019). Permission to feel. Celadon Books. https://us.macmillan.com/books/9781250212847/permissiontofeel

Duckworth, A. L. (2016). Grit: The power of passion and perseverance. Scribner. https://www.simonandschuster.com/books/Grit/Angela-Duckworth/9781501111112

Dweck, C. S. (2006). Mindset: The new psychology of success. Random House. https://www.penguinrandomhouse.com/books/44330/mindset-by-carol-s-dweck-phd/

Eisenberg, N., Fabes, R. A., & Spinrad, T. L. (2006). Prosocial development. In Handbook of child psychology (Vol. 3, pp. 646–718). Wiley. https://doi.org/10.1002/9780470147658.chpsy0311

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development. Child Development, 72(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience. Educational Psychologist, 47(4), 302–314. https://doi.org/10.1080/00461520.2012.722805

CONSTRUCT → OUTCOMES

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis. Child Development, 82(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Good, R. H., & Kaminski, R. A. (2002). Dynamic Indicators of Basic Early Literacy Skills (DIBELS), 6th ed. Institute for the Development of Educational Achievement. https://dibels.uoregon.edu/

Mahoney, J. L., Durlak, J. A., & Weissberg, R. P. (2018). An update on social and emotional learning outcome research. Phi Delta Kappan, 100(4), 18–23. https://doi.org/10.1177/0031721718815668

OECD. (2021). Beyond academic learning: First results from the survey of social and emotional skills. OECD Publishing. https://doi.org/10.1787/92a11084-en

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., et al. (2019). A national experiment reveals where a growth mindset improves achievement. Nature, 573, 364–369. https://doi.org/10.1038/s41586-019-1466-y

AUTOMATED SCORING METHODOLOGY

ETS. (2018). ETS guidelines for automated scoring. Educational Testing Service.

Ober, T. M., Zhang, S., Zapata-Rivera, D., Schroeder, N., & Botelho, A. (2026). Using LLMs to identify indicators of persistence from students' dialogues with a pedagogical agent. Journal of Educational Data Mining, 18(1), 208–243. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/1007

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

SELF-REPORT MEASUREMENT LIMITATIONS

EKrosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5(3), 213–236. https://doi.org/10.1002/acp.2350050305

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 224–239). Guilford. https://psycnet.apa.org/record/2007-04733-013

IMPACTER DOCUMENTATION

IMPACTER Pathway. (2025). Model Card v6.0: CORAL Ordinal Regression Models.

IMPACTER Pathway. (2025). How IMPACTER Measures Human Skills: The Linguistic Architecture of Growth.

IMPACTER Pathway. (2024). Anchor Attributes and Their Measurable Outcomes at School.