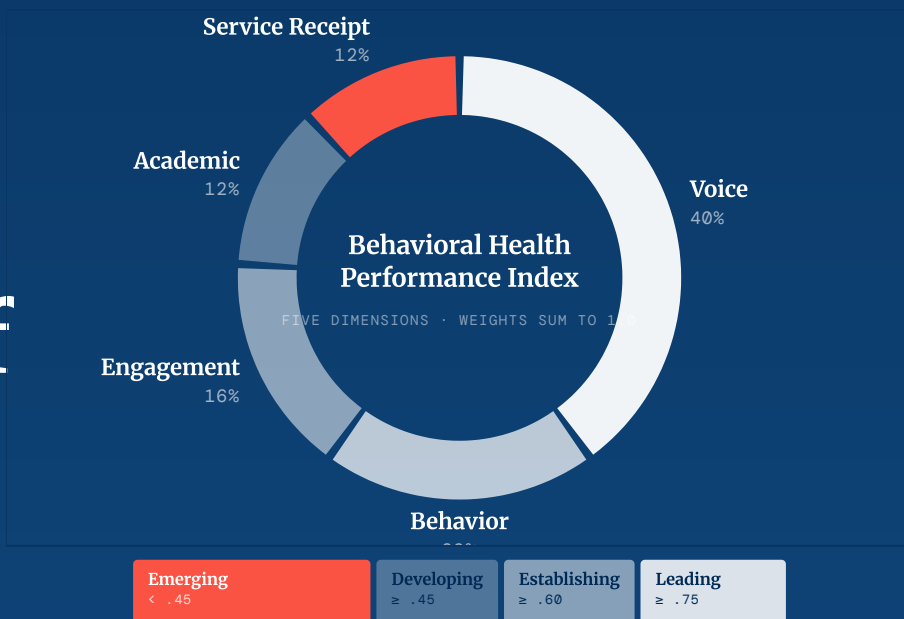


DELIVERABLE 2 · IMPLEMENTATION WORKPLAN CROSSWALK
ACTIVITIES 2.1 TO 2.7

Indicators, Outcomes & Metrics Brief

*What The System Counts, And How
It Knows The Counts Hold.*

The indicators, outcomes and metrics
the program reports to the Commission
across its four reporting domains, and
the methodology behind the Behavioral
Health Performance Index.



4

REPORTING DOMAINS

13

CANDIDATE INDICATORS

7

LITERATURE SECTIONS

66

SOURCES IN THE REGISTRY

PREPARED FOR

Behavioral Health Services Oversight and Accountability Commission

CONTRACT

Standard Agreement #25BHS0AC019

PREPARED BY

Santa Clara County Office of Education, with Impacter Pathway and Stanford

DUE TO THE COMMISSION

October 30, 2026 · Final November 30, 2026

LED BY THE SANTA CLARA COUNTY OFFICE OF EDUCATION

CONTENTS

What Is In This Brief

One sheet per section of the record: the section's own closing summary beside a figure built from its data. The full prose, with every source beside it, is the reader at bhpm.care/deliverables/d2.

00	What Deliverable 2 Answers		3
01	The Screener Landscape: What Schools Measure Today	Activity 2.1	5
02	From Self-Report To Authentic Voice: Language As Measurement Signal	Activity 2.1	10
03	Trustworthy Automated Scoring: The Evaluation Standards	Activity 2.1	14
04	Process And Outcome Indicators In MTSS-Aligned Behavioral Health	Activity 2.1 with Activity 2.3	18
05	Implementation Science And Process Evaluation	Activity 2.1	23
06	Composite Index Construction	Activity 2.1	27
07	The Impacter Pathway Research Lineage	Activity 2.1	30
08	The Behavioral Health Performance Index · Construction Methodology	Activity 2.4	34
09	Stakeholder Engagement And The Local-Framework Crosswalk	Activity 2.2	41
10	The Metrics Brief · Version 2a	Activity 2.5	43
11	The Annotated Bibliography		48
12	The Core Documents Behind The Brief		54

COMPILED

July To August 2026

Version 2a

DUE TO THE COMMISSION

October 30, 2026

Submission

FINAL

November 30, 2026

After The Activity 2.6 Review

How To Read This Brief

Each section sheet opens with the workplan activity it answers, quoted from the [Implementation Workplan Crosswalk](#), and closes on the section's own Bottom Line. Every figure traces to the record's data tables, the source registry or the earlier working design; nothing on these pages is stated for the first time here.

The Record Behind It

The twelve sections are published in full at bhpm.care/deliverables/d2, each with the sources it cites beside the prose, and the [earlier working design](#) keeps the interactive evidence tables and diagrams. This brief is the summary; the reader is the record.

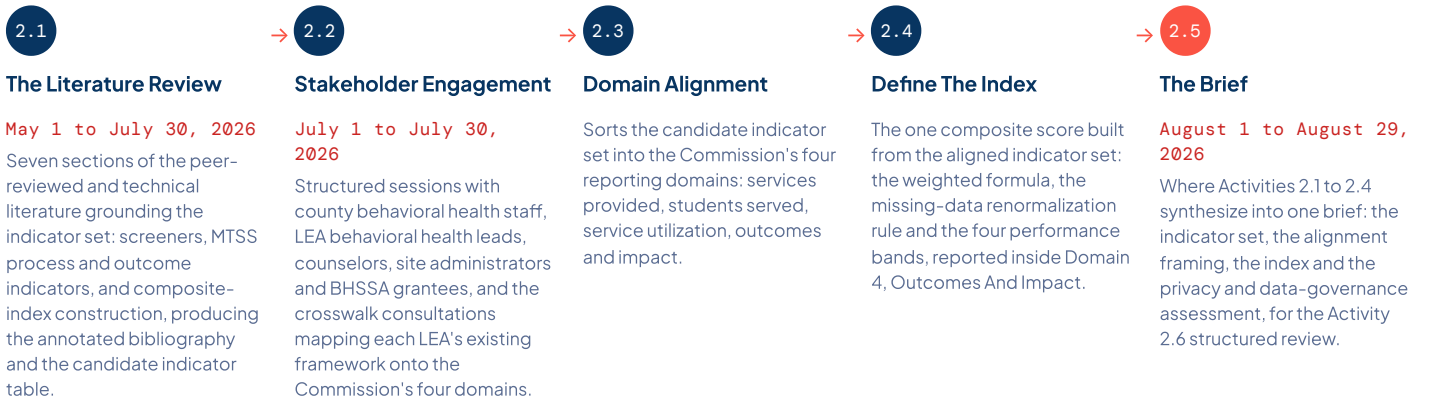
OVERVIEW · THE ACTIVITIES

What Deliverable 2 Answers

Deliverable 2 answers one question for the Commission: what should a school behavioral health performance management system count, and how does it know the counts are trustworthy. It proposes an indicator set, organizes that set under the Commission's four reporting domains, and defines the one composite score reported as the program's flagship outcome measure: the Behavioral Health Performance Index.

4 Reporting Domains	13 Candidate Indicators	7 Literature Sections	66 Sources In The Registry
---	---	---	--

How The Indicator Set Was Identified



WHAT THE RECORD ESTABLISHES

The evidence base, the index methodology and the brief itself sit in one record, each traceable to the workplan activity that produced it.

WHAT ACTIVITY 2.6 SETTLES

Three things in this record are settled by the Activity 2.6 structured review with all six LEA teams. The index weighting, banding and cadence questions run through that review. The stakeholder engagement record carries the mapping pattern the consultations use, and Deveau Burr Group's session log carries the count and the per-LEA breakdown. And the service-utilization indicators appear as reporting categories with a feasibility flag rather than as figures, because each waits on billing or data-sharing infrastructure some sites do not have yet.

OVERVIEW · THE SEVEN ACTIVITIES

Activities 2.1 To 2.7, As The Crosswalk Scopes Them

ACTIVITY	WHAT IT PRODUCES	LEAD	WHERE IT STANDS
2.1 05/01/26–07/30/26	The literature review: seven sections and the annotated bibliography behind them.	Impacter Pathway · Stanford Medicine	Written · sections 01–07 below
2.2 07/01/26–07/30/26	Stakeholder sessions with LEA and county staff, and the crosswalk of each LEA's local framework onto the Commission's four domains.	Deveau Burr Group	Sessions run · component 09
2.3	The candidate indicator set sorted into the Commission's four reporting domains.	Impacter Pathway	Written · component 04
2.4	The Behavioral Health Performance Index: formula, bands and missing-data rule.	Impacter Pathway	Version 0.9 · component 08
2.5 08/01/26–08/29/26	The Metrics Brief itself, synthesized from 2.1 through 2.4.	Impacter Pathway	Version 2a · component 10
2.6 from 09/01/26	LEA structured review, then submission to the Commission.	SCCOE	From 1 September 2026
2.7	Project management and administration across Activities 2.1–2.6.	SCCOE	Ongoing · no document
The seven Deliverable 2 activities as scoped in the Implementation Workplan Crosswalk.			

The One Composite Score The Brief Reports

The Behavioral Health Performance Index, version 0.9: five weighted dimensions summing to 1.0, reported inside Domain 4, Outcomes And Impact. Weights and bands are from the index methodology; they finalize through the Activity 2.6 structured review.



The Candidate Indicators, Counted By Domain

4 Services Provided	3 Students Served	3 Service Utilization	3 Outcomes And Impact	13 Candidate Indicators In Total
------------------------	----------------------	--------------------------	--------------------------	-------------------------------------

MISSING DATA

If a site is missing data for one of the five dimensions, that dimension is excluded and the remaining weights scale up proportionally so they still sum to 1.0. A site is never penalized for a data gap, it is scored on what it has actually reported.

BENEATH VOICE

Every student needs at least 4 of the 6 competencies reported to receive a Wellness Index score. If one or two are unreported, the index renormalizes across the remaining reported competencies, the same proportional-scaling logic BHPI itself uses at the site level.

SECTION 01 • THE SCREENER LANDSCAPE

The Screener Landscape: What Schools Measure Today

SCOPE • IMPLEMENTATION WORKPLAN CROSSWALK

Implementation Workplan Crosswalk Activity 2.1: the literature review grounding the indicator set. This section surveys the screening instruments districts already run and the structural gaps between them.

The Eight Instruments, By Who Reports And What The Score Can Trigger

From the section's comparison table. A filled mark is an informant the instrument asks; the actionability column is the section's own reading of each.

INSTRUMENT	STUDENT	TEACHER	PARENT	CLINICIAN	FORMAT	INDIVIDUAL ACTIONABILITY
PHQ-9	●	•	•	•	9-item Likert, past 2 weeks, DSM-IV-mapped depression severity	Individual score plus diagnostic-algorithm cutoff; built for direct clinical follow-up
GAD-7	●	•	•	•	7-item Likert, past 2 weeks, generalized anxiety severity	Individual score and cutoff; companion instrument to PHQ-9 for clinical triage
BASC-3 BESS	●	●	•	•	Brief multi-item behavioral/emotional risk rating; higher-order factor structure (Teacher Form), bifactor structure (Student Form)	Widely adopted for individual risk tiering, but the predecessor BASC-2-BESS teacher form shows meaningfully weaker predictive accuracy for internalizing disorders (AUC .58) than externalizing (AUC .73) against clinician diagnosis
SAEBRS	•	●	•	•	12-item multi-domain (social, academic, emotional) risk screener, developed from the earlier SABRS	A multiple-gating procedure combining teacher nomination with SAEBSRS scores supports tiered individual identification
SDQ	●	●	●	•	25-item, five-scale instrument (emotional symptoms, conduct problems, hyperactivity/inattention, peer problems, prosocial behavior)	Teacher-rated SDQ shows the same predictive-accuracy pattern as BASC-2-BESS against clinical diagnosis: AUC .73 for externalizing disorders vs. .58 for internalizing disorders
SRSS-IE	•	●	•	•	Brief dual-scale risk screener covering both internalizing and externalizing risk	Dual-dimension score supports differentiated tiering; predictive validity evidence is strongest at the elementary level
C-SSRS	●	•	•	●	Semi-structured interview assessing suicidal ideation and behavior severity	An individual, crisis-tier instrument built for direct safety-planning response, not universal classroom screening
CHKS / CalSCHLS	●	•	•	•	Anonymous statewide survey with a standardized Core Module (required in grades 7 and 9)	None at the individual level by design: anonymized, aggregate-only reporting

WHAT THE SECTION COVERS

- Adult-Clinical Screeners Adapted For Schools
- Instruments Built For Schools
- Four Structural Gaps Beneath The Landscape
- The Screening-To-Service Gap
- Comparing The Eight Instruments

The full section, with every source beside the prose: bhpm.care/deliverables/d2/screener-landscape

BOTTOM LINE • THE RECORD'S OWN SUMMARY

BHPM's indicator set holds a deliberate position inside this landscape: a complement to the eight instruments surveyed above, not a replacement for any of them. That position starts with the anonymity-actionability line running through the landscape: CHKS at one pole, the C-SSRS at the other, and the six school-based universal screeners falling at varying degrees of individual actionability between them. Naming that line explicitly, and showing where an indicator sits on it, is what lets a reader see which tradeoff is being made rather than discover it by implication. The PLAY-MH internalizing/externalizing accuracy gap (AUC .73 vs. .58), reinforced by Siceloff and colleagues' district-level finding that universal screeners under-detect internalizing risk in practice as well as in controlled comparison, is the evidence that grounds the case for language-based measurement in Section 2: a concrete detection failure, not a general appeal to innovation.

SECTION 01 · THE SCREENER LANDSCAPE · CONTINUED

The Screener Landscape: What Schools Measure Today

What should a school behavioral health performance management system count, and how does it know the counts are trustworthy? Before answering either question, it helps to know what schools already count. Districts entering behavioral health screening do not start from a blank page: a substantial instrument landscape already exists, built across four decades of clinical and school psychology research, and BHPM's indicator design has to sit deliberately inside that landscape rather than around it. That landscape did not develop as one coherent design; it accumulated in three distinct waves, each answering a different question. Adult clinical medicine produced checklist screeners meant to trigger a diagnosis. School psychology, working a decade or two later and often against those same clinical instruments as a baseline, built its own tools meant to trigger a tiered-support decision rather than a clinical one. And state infrastructure built a third kind of instrument meant to describe a population rather than flag any individual student at all. Treating these as interchangeable rows in one comparison table, rather than as three different tools built to answer three different questions, misreads what each one can defensibly be asked to do. The eight instruments surveyed here also split along a second, orthogonal axis worth naming up front: who is asked to report. Most are self-report or teacher-report checklists resting on a single informant's judgment; BASC-3 BESS and the SDQ each gather more than one informant's perspective on the same student rather than relying on just one; the Columbia-Suicide Severity Rating Scale substitutes a clinician's structured interview for either; and CHKS is answered by the student but never resolves to an individual score at all. None relies on a performance task or an observed-behavior sample independent of someone's rating, a gap that matters later in this section, because it means every accuracy figure quoted below is an accuracy figure about how well one kind of report predicts another, not about how well any of them capture behavior directly.

Adult-Clinical Screeners Adapted For Schools

Two of the most widely used instruments in that landscape were never built for schools at all. The PHQ-9 [\(KROENKE ET AL. 2001\)](#) and the GAD-7 [\(SPITZER ET AL. 2006\)](#) are adult-primary-care screeners, validated against structured clinical interviews and later adapted downward into adolescent and school-based use because nothing purpose-built displaced them. Both are self-report checklists mapped directly onto DSM criteria, both return a single cutoff score with a published sensitivity/specificity profile, and both were designed to trigger one specific downstream action: a clinical referral decision. The PHQ-9's original validation, run on 6,000 primary-care and ob-

gyn patients with a criterion sample of 580 assessed by structured interview, found that a cutoff of 10 or above identified major depression with 88 percent sensitivity and 88 percent specificity, and the same study found severity climbing in step with worse functional status across every SF-20 subscale and with higher sick-day and healthcare-utilization rates, evidence that the instrument works as both a categorical diagnostic algorithm and a continuous severity scale, not only the former. The GAD-7 was validated the same way for generalized anxiety, with a cutoff of 10 yielding 89 percent sensitivity and 82 percent specificity, and its validation work established the now-standard "past two weeks" self-report window and confirmed good internal consistency alongside factorial, criterion, and procedural validity across a large primary-care cohort. Their durability across two decades of behavioral health practice is real, but so is the fact that they were built to answer a narrower question than a school system asks: not "how is this student doing across the domains a district's behavioral health framework cares about," but "does this specific respondent meet criteria for this specific disorder today."

Instruments Built For Schools

The instruments built for schools take a different shape. BASC-3's Behavioral and Emotional Screening System (BESS) is the most commercially entrenched of them, with a Teacher Form whose higher-order factor structure holds up under confirmatory analysis on a K-5 sample of 1,472 students and shows configural, metric, and scalar invariance across student age [\(EKLUND ET AL. 2022\)](#), invariance solid enough that the same study found BESS scores predicting academic achievement, behavioral engagement, and school climate outcomes, not just concurrent risk status. A Student Form extends that validity evidence to self-report: in a sample of 737 urban students in grades 4 through 8, its bifactor structure held and measurement equivalence was demonstrated specifically across White, Black, and Latinx students and across gender [\(BASTING ET AL. 2022\)](#), a fairness finding that matters directly to whether a commercial screener's psychometrics travel across a district's own student population rather than only across the sample it was validated on. SAEBRS took a similar structural-validity path from a different starting point. It grew out of an initial 12-item Social and Academic Behavior Risk Screener built through a content-validation process plus reliability and exploratory factor analysis on 243 students rated by 54 teachers, with an item pool drawn deliberately from developmental research on behavior-problem trajectories and social and academic competence models rather than from an existing clinical instrument [\(KILGUS ET AL. 2013\)](#). From there it added a multiple-gating procedure that combines teacher nomination with screener scores, and a two-study investigation anchored by an 864-student sample demonstrated acceptable-to-optimal internal consistency, concurrent validity, and diagnostic accuracy across elementary and middle school settings [\(KILGUS ET AL. 2016\)](#).

SECTION 01 · THE SCREENER LANDSCAPE · CONTINUED

Instruments Built For Schools

It was later subjected to an argument-based validation using bifactor item response theory and structural equation modeling, a bifactor model that outperformed unidimensional and correlated-factors alternatives, discriminated moderate- from high-risk students, held up under a hierarchical-omega reliability check on its Total Behavior score, and was tested for convergent and discriminant relations directly against the BASC BESS itself (KILGUS ET AL. 2018), a maturity benchmark few school screeners reach and one that effectively cross-validates two of this landscape's most entrenched instruments against each other rather than asking either to stand alone. Both instruments carry that commercial, gated-procedure weight forward into the cost-and-time comparison below, a dimension the psychometric literature above does not measure at all, but one that shapes which districts can realistically field either instrument at scale.

The Strengths and Difficulties Questionnaire took the opposite path to the same destination. Rather than commercial packaging or a gated procedure, its 25-item, five-scale design was tested in a clinical sample of 403 children against the established Rutter questionnaires, and a five-factor structure, four difficulty scales plus a prosocial-behavior scale, replicated across parent and teacher raters, not just one informant type (GOODMAN 1997). That multi-informant replication, combined with free and noncommercial distribution, made the SDQ the most internationally adopted screener in this landscape. SRSS-IE occupies similar territory domestically: a free, brief, dual-scale teacher rating whose internalizing and externalizing risk classifications predicted subsequent academic and behavioral outcomes across 4,465 elementary students in 14 schools (LANE ET AL. 2019), with the dual structure itself shown to carry independent predictive value over a single-dimension screener, evidence that splitting risk into two channels rather than blending it into one score is not just conceptually tidier, it captures something a single score misses.

Two more instruments sit at the edges of this landscape rather than inside its center, and both matter to how BHPM should think about scope. The Columbia-Suicide Severity Rating Scale is not a universal screener at all; it is a semi-structured, clinician-administered interview built for one purpose, distinguishing degrees of suicide risk severity, and its validation across three multisite adolescent and adult clinical trials established convergent and divergent validity against existing suicidality measures and found that lifetime worst-point ideation with intent or a plan predicted roughly four times the future attempt risk of current ideation alone (POSNER ET AL. 2011), a finding robust enough across three independent samples that the instrument became the standard reference not only for clinical trials but for school crisis-response protocols specifically. It belongs at the individual crisis-response tier, not the universal-screening tier, and a metrics framework that conflates the two risks either over-medicalizing a screening program or under-resourcing a crisis one. At the opposite end sits California's own state-sponsored infrastructure, the California Healthy Kids Survey within the CalSCHLS system: a standardized, anonymous survey administered by WestEd

for the California Department of Education since 1997, developed together with Duerr Evaluation Resources and expert advisory committees, with a Core Module required in grades 7 and 9 and current participation above 69 percent of California districts. CHKS is deliberately built to produce no individual-level score at all (WESTED. 2026). That design choice, anonymity purchased at the cost of individual actionability, is not a limitation the instrument's designers overlooked; it is the tradeoff the rest of this section turns on.

Four Structural Gaps Beneath The Landscape

Four structural gaps run underneath this entire landscape, and each one bears directly on what a BHPM indicator set has to be built to withstand. The first is a method-effect problem inherent to self-report and teacher-report measurement itself. Duckworth and Yeager's review of how personal qualities beyond cognitive ability get measured, using self-control as its worked case study, found that no single method (self-report, other-report, or performance task) is free of bias, and that terminology debates over what to call these qualities obscure far more agreement than disagreement about which specific qualities actually matter (DUCKWORTH ET AL. 2015). Their conclusion was not to identify the least-biased method and standardize on it; every method they examined carried some version of the same limitation, which is why they recommended triangulating across methods rather than trusting any single one at high-stakes scale. Every instrument named above inherits some version of this limitation regardless of how well its factor structure replicates, and the informant split named at the top of this section is exactly where that limitation lives: a teacher-report instrument like SAEBRS or the BESS Teacher Form is vulnerable to whatever a teacher can and cannot observe, a self-report instrument like the PHQ-9 or GAD-7 is vulnerable to whatever a student is willing to disclose on a checklist, and even the two instruments that gather more than one informant's view, BASC-3 BESS and the SDQ, report each informant's rating separately rather than combining them into a single score that could offset either one's blind spot.

The second is the anonymity-actionability tradeoff CHKS makes explicit but does not resolve. A screening instrument can be built to protect a respondent's identity or built to trigger an individual follow-up action; the instruments surveyed here sit at different points along that line (CHKS at one pole, C-SSRS at the other, and the six school-based universal screeners in between at varying degrees of individual actionability), and a district cannot simply average across them without deciding, deliberately, where its own program needs to sit. That decision is entangled with the informant-type split described above rather than separate from it: an individually actionable score requires knowing which student produced it, which in turn requires an identified informant, so a district cannot adopt CHKS's anonymity and still expect PHQ-9-style individual follow-up out of the same instrument. The tradeoff is a design constraint on the whole indicator set, not a property of any one screener that a district could simply screen for and avoid.

SECTION 01 · THE SCREENER LANDSCAPE · CONTINUED

Four Structural Gaps Beneath The Landscape

The third gap is acceptability, and it has its own growing evidence base, now with a resourcing dimension attached to it. Two systematic reviews from 2019 and 2020 found the evidence on school-based mental health identification mixed at best: feasibility and acceptability findings split along stakeholder lines, with parents, students, and health professionals generally supportive and school staff more skeptical, largely over role-scope concerns, though the same review noted that near-universal reach and proximity to a school's existing support resources are exactly what make screening worth the friction in the first place (SONESON ET AL. 2020), while a companion review found the effectiveness and cost-effectiveness evidence base for these programs still limited and methodologically uneven, and identified specific gaps in how that research has so far been designed (ANDERSON ET AL. 2019). A 2025 systematic review applying the Theoretical Framework of Acceptability specifically to universal mental health screening brought that evidence current, using the Mixed Methods Appraisal Tool to grade study quality and a narrative synthesis to separate acceptability across students, caregivers, and school staff as three distinct stakeholder groups with distinct drivers and barriers, rather than compressing all three into a single undifferentiated "acceptability" score (PALMER ET AL. 2025). A district-level implementation study adds a piece none of the three reviews measures directly: a CDC-funded district screening effort found that fewer than 15 percent of U.S. schools had, at that point, implemented any systematic behavioral-health screening procedure at all, and that even a funded, university-supported rollout reached only 68.6 and 68.5 percent student completion across two successive years, dropping to 56 and 57 percent at the high school level specifically (SICELOFF ET AL. 2017). Acceptability, in other words, is not only a matter of whether stakeholders approve of screening in principle; it is also a matter of whether a district without grant-funded technology and data-management support can actually get the screener into most students' hands in the first place.

The Screening-To-Service Gap

The fourth gap is the one the psychometric literature states most concretely, and it is a screening-to-service gap hiding inside an aggregate accuracy number. The CDC's multi-site Project to Learn About Youth-Mental Health dataset tested how well teacher-completed BASC-2-BESS and SDQ scores, gathered across four districts, predicted parent-report-confirmed diagnoses assessed through a structured clinical interview, in a sample of 1,054 K-12

students: both instruments reached an AUC of .73 for externalizing disorders but only .58 for internalizing ones, leading the study's own authors to conclude that both screeners carry only modest predictive utility against a rigorous diagnostic criterion (DANIELSON ET AL. 2023). A single reported accuracy figure for either screener would flatten that gap invisibly. The instruments in this landscape are, on average, reasonably good at flagging the students whose distress is visible to a teacher and considerably weaker at flagging the students whose distress is not, which is precisely the population a behavioral health performance management system exists to find. Sicheloff and colleagues' feasibility work reaches a similar conclusion from the implementation side rather than the psychometric one: the universal screeners in their district-wide sample showed stronger sensitivity for externalizing problems than for internalizing ones, which is part of why newer dual-scale instruments like SRSS-IE were built to separate the two risk channels rather than blend them into a single score (SICELOFF ET AL. 2017). Two independent evidence bases, one measuring diagnostic accuracy in a controlled comparison and the other measuring what happens in an actual district rollout, converge on the same finding: internalizing risk is the harder detection problem here, not a minor addendum to externalizing risk.

Comparing The Eight Instruments

None of the four structural gaps above tracks cleanly onto cost, and that is itself worth stating plainly before the comparison table does the rest of the work. The free instruments in this landscape, the SDQ, SRSS-IE, and CHKS, are not free of the method-effect, acceptability, or screening-to-service problems that burden the commercial and clinical instruments beside them; they simply carry those same evidentiary limitations at a lower price and a lower staff-time cost. A district weighing BASC-3 BESS's stronger cross-validated structural evidence against SRSS-IE's lower cost is not trading rigor for savings in any simple sense. It is trading one already-imperfect method for another, at a different price point, inside the same set of structural constraints. The table below lays these eight instruments side by side across the dimensions that matter for indicator design, not just psychometric quality in isolation: who reports, in what format, with what accuracy evidence, whether that accuracy translates into an individually actionable score, and what it costs a district in dollars and staff time to run. Selecting each row opens its full source profile.

INSTRUMENT	INFORMANT	FORMAT	PSYCHOMETRICS	INDIVIDUAL ACTIONABILITY	COST & TIME
PHQ-9	Student self-report	9-item Likert, past 2 weeks, DSM-IV-mapped depression severity	≥10 cutoff: 88% sensitivity / 88% specificity against a structured clinical interview (n=580)	Individual score plus diagnostic-algorithm cutoff; built for direct clinical follow-up	Public domain; roughly 5 minutes to administer
GAD-7	Student self-report	7-item Likert, past 2 weeks, generalized anxiety severity	Cutoff of 10: 89% sensitivity / 82% specificity in a large primary-care validation sample	Individual score and cutoff; companion instrument to PHQ-9 for clinical triage	Public domain; roughly 3-5 minutes to administer
BASC-3 BESS (Teacher and Student Forms)	Teacher rating (Teacher Form); student self-report (Student Form)	Brief multi-item behavioral/emotional risk rating; higher-order factor structure (Teacher Form), bifactor structure (Student Form)	Confirmed factor structure with measurement invariance across student age (Teacher Form) and across race/ethnicity and gender (Student Form)	Widely adopted for individual risk tiering, but the predecessor BASC-2-BESS teacher form shows meaningfully weaker predictive accuracy for internalizing disorders (AUC .58) than externalizing (AUC .73) against clinician diagnosis	Proprietary, publisher-licensed instrument; brief teacher or student rating

SECTION 01 · THE SCREENER LANDSCAPE · CONTINUED

Comparing The Eight Instruments

INSTRUMENT	INFORMANT	FORMAT	PSYCHOMETRICS	INDIVIDUAL ACTIONABILITY	COST & TIME
SAEBRS (Social, Academic, and Emotional Behavior Risk Screener)	Teacher rating	12-item multi-domain (social, academic, emotional) risk screener, developed from the earlier SABRS	Acceptable-to-optimal internal consistency and concurrent validity; bifactor IRT/SEM modeling supports score interpretation and discriminates moderate- from high-risk students	A multiple-gating procedure combining teacher nomination with SAEBSRS scores supports tiered individual identification	Brief teacher rating; the multiple-gating procedure adds a second screening step
Strengths and Difficulties Questionnaire (SDQ)	Multi-informant: student, parent, and teacher versions	25-item, five-scale instrument (emotional symptoms, conduct problems, hyperactivity/inattention, peer problems, prosocial behavior)	Strong correspondence with the established Rutter questionnaires; five-factor structure replicated across raters	Teacher-rated SDQ shows the same predictive-accuracy pattern as BASC-2-BESS against clinical diagnosis: AUC .73 for externalizing disorders vs. .58 for internalizing disorders	Free, noncommercial distribution; a major driver of its international adoption
SRSS-IE (Student Risk Screening Scale — Internalizing and Externalizing)	Teacher rating	Brief dual-scale risk screener covering both internalizing and externalizing risk	Risk classifications predicted subsequent academic and behavioral outcomes in a sample of 4,465 elementary students	Dual-dimension score supports differentiated tiering; predictive validity evidence is strongest at the elementary level	Free; brief teacher rating designed for large-scale universal screening
Columbia-Suicide Severity Rating Scale (C-SSRS)	Student, via trained-clinician-administered interview	Semi-structured interview assessing suicidal ideation and behavior severity	Lifetime worst-point ideation with intent or plan predicts roughly 4x the future attempt risk of current ideation alone	An individual, crisis-tier instrument built for direct safety-planning response, not universal classroom screening	Requires a trained administrator; not a self-administered classroom tool
California Healthy Kids Survey (CHKS), within the CalSCHLS system	Student self-report	Anonymous statewide survey with a standardized Core Module (required in grades 7 and 9)	State-sponsored, standardized administration in place since 1997, currently used by more than 69% of California districts	None at the individual level by design; anonymized, aggregate-only reporting	State-sponsored infrastructure; no per-student licensing cost to districts
Eight instruments compared across informant, format, psychometric evidence, individual actionability and cost. Row detail and per-row sources are in the earlier working design.					

SOURCES CITED IN THIS SECTION · 17

Anderson et al., 2019 The evidence base on effectiveness and cost-effectiveness of school-based identification remains limited and heterogeneous.	Basting et al., 2022 A bifactor structure best fit BESS Student Form data among urban middle-grades students.
Danielson et al., 2023 Both BASC-2-BESS and SDQ teacher forms showed AUC=.73 for predicting any externalizing disorder.	Duckworth et al., 2015 No single measurement method (self-report, other-report, performance task) is free of bias; each carries distinct method effects.
Eklund et al., 2022 A higher-order factor structure provided acceptable fit to BASC-3 BESS Teacher Form data.	Goodman, 1997 SDQ scales showed strong correspondence with the established Rutter questionnaires.
Kilgus et al., 2013 A content-validation process plus reliability and exploratory factor analysis yielded a 12-item screener.	Kilgus et al., 2016 SAEBRS-TRS showed acceptable to optimal internal consistency, concurrent validity, and diagnostic accuracy across elementary and middle school samples.
Kilgus et al., 2018 A bifactor IRT model adequately fit SAEBSRS items, discriminating moderate- from high-risk students.	Kroenke et al., 2001 A PHQ-9 cutoff of ≥10 yielded 88% sensitivity and 88% specificity for major depression against a structured interview criterion.
Lane et al., 2019 SRSS-IE risk classifications predicted subsequent academic and behavioral outcomes at the elementary level.	Palmer et al., 2025 The first systematic review to apply the TFA specifically to school-based UMHS acceptability.
Posner et al., 2011 Lifetime worst-point suicidal ideation with intent (or intent with plan) predicted future suicide attempt risk roughly 4x that of current ideation alone.	Siceloff et al., 2017 Fewer than 15% of U.S. schools had, at the time of writing, implemented systematic procedures to evaluate students' behavioral/emotional health needs district-wide.
Soneson et al., 2020 Feasibility/acceptability evidence for school-based screening is mixed overall.	Spitzer et al., 2006 A cutoff of 10 yielded 89% sensitivity and 82% specificity for generalized anxiety disorder.
WestEd., 2026 Anonymized, aggregate-only data collection model with no individual-student-level identification.	

SECTION 02 · FROM SELF-REPORT TO AUTHENTIC VOICE

From Self-Report To Authentic Voice: Language As Measurement Signal

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.1: why open-response student language is a measurement signal distinct from checklist self-report, and what the published research does and does not support.

WHAT A CHECKLIST ITEM COLLECTS

☐ Not at all

☐ Several days

☐ More than half the days

One number per item, the same whether the student is guarded, articulate, distracted or translating the item in their head. Lexical choice, structure, elaboration, hedging and framing are discarded.

LANGUAGE-BASED DETECTION · POOLED FIGURES, FISHER AND COLLEAGUES 2026



Shared 0 to 1 axis. The spread across metrics is the section's own point: the achievable number depends on which metric a study chose to report.

WHAT THE SECTION COVERS

- The Clinical Evidence Base For Language-Based Detection
- Risk Exposure And The Multilingual Question
- Educational Measurement's Own Track Record
- Construct Clarity, Not Model Scale, Drives Reliability
- Impacter Pathway's Discriminative Scoring Architecture

The full section, with every source beside the prose: bhpm.care/deliverables/d2/authentic-voice

BOTTOM LINE · THE RECORD'S OWN SUMMARY

Language-based measurement is a real but bounded research area (Fisher et al., 2026; Low et al., 2020), not an established clinical capability, and the bias, privacy, and multilingual risks that research documents (Teferra et al., 2024) are design constraints Impacter Pathway has engineered against, not risks discovered after deployment. ETS's conversation-based-assessment precedent (Zapata-Rivera & Forsyth, 2022; Lopez et al., 2021) is the closer analogue to a K-12 open-ended-response context than the clinical literature, and it is the stronger anchor for this measurement approach than the depression-detection numbers are.

The architectural case is the strongest one, and it carries forward into Section 3's evaluation-standards case: Ober et al. (2026) found that construct-definition clarity, not model scale, drives reliable coding, and documented real reproducibility problems in generative LLM-based coding, reinforced by Turpin et al. (2023): a conclusion validity theory reaches independently as well (Bauer & Zapata-Rivera, 2020). Impacter Pathway's discriminative encoder-plus-CORAL ordinal pipeline (Model Suite v6.0, DeBERTa-v3-base, 184M parameters, frozen and versioned) is one reasonable structural response to that finding: a design choice this evidence supports, not a claim this evidence proves, and a categorically different architecture than a generative model reasoning its way to a score.

The scope stays disciplined by design: rubric-aligned ordinal scoring, never phrased as a clinical or diagnostic claim.

SECTION 02 · FROM SELF-REPORT TO AUTHENTIC VOICE · CONTINUED

From Self-Report To Authentic Voice: Language As Measurement Signal

A behavioral health screener asks a student to rate a checklist item against a shared scale ("not at all," "several days," "more than half the days") and treats the resulting number as the measurement. The instrument does the same work whether the student circling "several days" is guarded, articulate, distracted, or silently translating the item from another language in their head; the scale absorbs all of that into one number and discards the rest. Unconstrained language works differently. When a student answers an open-ended prompt in their own words, the response carries lexical choice, structure, elaboration, hedging, and framing that a five-point item cannot capture, and a growing body of clinical and educational measurement research treats that language itself as a scorable signal rather than as color commentary around the "real" score. This section lays out what that research base currently supports, what it does not, and why Impacter Pathway's production scoring architecture is built the way it is in response to it.

The Clinical Evidence Base For Language-Based Detection

The clinical evidence base for language as a measurement signal is largest around depression detection. A 2026 systematic review and meta-analysis in *npj Digital Medicine* screened 892 records, included 123 studies covering 129 effects, and pooled results across 50 independent datasets to estimate what natural-language and machine learning methods can currently achieve when detecting depression from spoken or written language (FISHER ET AL. 2026). The pooled estimate across 43 studies and 40,983 text samples was 0.80 accuracy (95% CI 0.76–0.83), with precision of 0.78 across 28

studies, recall of 0.76 across 33 studies, an AUC of 0.79 across 14 studies, and a more conservative balanced-accuracy figure of 0.71 across the 16 studies that reported it, a spread across metrics that itself signals how much the achievable number depends on which metric a given study chose to report. The review's subgroup analysis, run by language, text source, feature type, and classifier, found accuracy was highest specifically in studies built on structured clinical interviews rather than casual or social-media text, in non-English-language samples rather than English-language ones, and in studies using linguistic or embedding-based features rather than surface word counts, a pattern on the language axis that argues the signal lives in how carefully something is elicited, not simply in which language it happens to be said in. A broader review the same year extended the picture beyond depression to automated speech-based assessment across schizophrenia, bipolar disorder, PTSD, anxiety, and eating disorders, screening 1,395 studies and including 127 (LOW ET AL. 2020). Most of that included literature still clustered around depression, schizophrenia, and bipolar disorder specifically; PTSD, anxiety, and eating disorders, despite being explicitly in scope for the review, turned up far less evidence, which is exactly the pattern this section keeps returning to. Set side by side, these two reviews establish a real but bounded ceiling: language-based detection of a single, well-studied clinical construct, assessed under a structured elicitation format, performs meaningfully above chance, and the evidence thins quickly once the target construct, the population, or the elicitation format moves away from that well-trodden center. The diagram below sets that pooled, metric-dependent spread next to the single-number mechanism the opening paragraph described, so the contrast is a comparison rather than a claim to take on faith.

Measurement comparison

A side-by-side of checklist self-report and open-response language measurement. It is drawn as a diagram in the earlier working design.

[Open measurement comparison in the archive](#)

Risk Exposure And The Multilingual Question

That evidence base also documents its own exposure. A 2024 review in the JMIR-affiliated *Interactive Journal of Medical Research* surveying natural-language-processing approaches to depression screening names the specific risks a system built on this method has to design against (TEFERRA ET AL. 2024). It flags reidentification risk in clinical text even after standard anonymization and recommends differential privacy as a companion safeguard rather than treating anonymization alone as sufficient. It documents bias in common word-embedding models, GloVe and Word2Vec by name, along religion, race, gender, nationality, sexuality, and age lines that can degrade accuracy specifically for underrepresented groups rather than uniformly across a sample. It names interpretability limits inherent to black-box deep learning as a validity concern in its own right, not just a usability inconvenience. And it calls explicitly for culturally and linguistically sensitive models, citing evidence that how distress gets expressed in language varies across languages and cultures. That last point lands directly on a design constraint BHPM

cannot defer: students will answer BHPM prompts in whatever language they think in, and a model trained mostly on English clinical text is exactly the kind of system this review is warning about. None of this is a reason to treat language-based measurement as unusable: the clinical NLP field would not be publishing systematic reviews of its own aggregate performance if the answer were simply "don't." It is a reason to treat every one of those failure modes (reidentification, embedding bias, opacity, and monolingual training data) as a design constraint to be engineered against from the start, not a problem discovered after deployment.

The multilingual angle threads through several of these findings, not just Teferra et al.'s caution. Fisher et al.'s own subgroup analysis found accuracy highest in non-English-language samples, not lowest, a result that runs against the intuition that English-trained pipelines transfer best, and suggests the studies driving that subgroup were built around structured, purpose-built elicitation in the target language rather than a repurposed English-first pipeline translated after the fact.

SECTION 02 · FROM SELF-REPORT TO AUTHENTIC VOICE · CONTINUED

Risk Exposure And The Multilingual Question

Lopez et al. and Ober et al., discussed below, sit on the same side of that finding from the education-measurement side: both evaluate systems interpreting language from students who are actively acquiring English, not fluent adult clinical patients, and both report the system performing well on that population specifically. None of that licenses treating multilingual administration as a solved problem, Teferra et al.'s call for culturally and linguistically sensitive models exists precisely because the underlying embedding models still carry documented bias, and one positive subgroup finding in one meta-analysis is not a validated multilingual pipeline. But it does mean the research base points toward purpose-built, in-language elicitation and scoring as the direction with actual evidentiary support behind it, not away from it, which matters directly for a screening population that will answer BHPM prompts in whatever language they think in.

Educational Measurement's Own Track Record

Educational measurement has its own, more directly relevant track record with open-ended student language, and it predates the current wave of clinical NLP research. ETS's conversation-based assessment program builds learner models that combine evidence from multiple-choice items, constructed responses, and dialogue with a conversational agent into a single account of what a student knows or can do, with the conversation's complexity adapting dynamically as the system updates its read on the student's real-time progress (ZAPATA-RIVERA ET AL. 2022). That adaptive design matters here because it treats open-ended dialogue as a live measurement channel, not a static text sample scored after the fact once the conversation is over. A 2021 ETS research report evaluated one such prototype, ELLA-Math, specifically on its ability to interpret middle-school English learners' written responses to math prompts, and found the system highly accurate at correctly classifying the sentences those students actually produced (LOPEZ ET AL. 2021). The scope of that finding is worth stating precisely: it is one prototype, evaluated in one subject domain, with one population, a real limitation, not a broad validation of automated interpretation across every context. But the population it was tested on is the relevant one. It is not a claim about fluent adult English text; it is a claim about interpreting open-ended written language from students who are actively acquiring the language they are writing in, which is closer to the linguistic diversity a multilingual, K-12 student population will actually produce than anything in the depression-detection literature above. That result matters specifically because it is a precedent for automated interpretation of open-ended language from students across a range of linguistic backgrounds, in an assessment context, rather than a claim about clinical text from adult patients.

Construct Clarity, Not Model Scale, Drives Reliability

The most directly load-bearing study for how Impacter Pathway's pipeline is built, though, is a 2026 study in the *Journal of Educational Data Mining* that compared several large language model configurations against expert human coders on more than 10,000 conversational turns from 107 English learners, coding constructs

including persistence, self-efficacy, interest, perceived challenge, and prior knowledge from a conversation-based math assessment (OBER ET AL. 2026). The study varied both model temperature and model size across multiple ChatGPT-4o configurations and compared each configuration's coding against expert human coders construct by construct, rather than reporting one blended agreement figure across all five constructs at once. The finding was not that a larger or more capable model closes the gap with human coders. It was that how precisely a construct is operationally defined governs human-model agreement far more than model scale does: self-efficacy, the most tightly and operationally defined construct in the study, showed the strongest human-model alignment and needed a low, near-deterministic temperature setting to hold that alignment steady, while the more theoretically diffuse constructs behaved inconsistently across different temperature settings, the same construct, coded by the same model family, produced different agreement levels depending on a decoding parameter that has nothing to do with the construct itself. That is a documented reproducibility problem, not a one-off measurement artifact, and it is open-access research, JEDM publishes without a paywall, with a supplemental materials record on OSF, which is part of why it is verifiable rather than taken on faith. A separate 2023 study accepted at NeurIPS reinforces that same instability from a different angle. It shows that large language models' own chain-of-thought explanations can misrepresent the actual basis for an output: when researchers quietly injected a biasing feature into the input (for instance, reordering which multiple-choice option came first), models still produced fluent, plausible-sounding reasoning that omitted the injected feature entirely, even though that feature measurably changed the model's answer, with accuracy dropping by as much as 36 percent across 13 benchmark tasks once the bias was in place (TURPIN ET AL. 2023). Set against each other, these two studies describe a specific, well-documented failure mode of generative language models used for construct coding: reasoning and consistency that look plausible turn by turn but are not reliably faithful to the same construct, in the same way, twice.

That construct-clarity argument is not new to Ober and colleagues; it echoes a long-standing position in validity theory. A chapter on the cognitive foundations of automated scoring makes the same case from first principles rather than from an empirical LLM comparison: scoring-system design should be traceable to an explicit cognitive model of the construct the assessment is meant to elicit, not just statistical fit to a set of human ratings, and the transparency of a scoring system's logic is itself a validity concern rather than a mere usability nicety (BAUER ET AL. 2020). Held up against Ober et al.'s empirical finding, the two arguments converge on the same conclusion from opposite directions (one from measurement theory, one from an LLM-coding experiment) that a scoring system's trustworthiness rests on how explicitly its target construct is defined and how transparently its logic can be traced back to that definition, not on how large or fluent the underlying model is.

SECTION 02 · FROM SELF-REPORT TO AUTHENTIC VOICE

Impacter Pathway's Discriminative Scoring Architecture

Ober and colleagues' own prescription is rubric-first construct definition, and Impacter Pathway's production architecture is built to satisfy that prescription structurally rather than through prompt engineering after the fact. The scoring pipeline is discriminative, not generative: a fine-tuned transformer encoder is paired with a CORAL ordinal regression head. The encoder descends from the disentangled-attention family introduced by He, Liu, Gao, and Chen, which represents each token with separate content and relative-position vectors and adds an enhanced mask decoder that incorporates absolute position information during pretraining, a design that, trained on half the data, still outperformed RoBERTa-Large on MNLI, SQuAD v2.0, and RACE (HE ET AL. 2021). A later refinement in that same lineage swaps in ELECTRA-style replaced-token detection and gradient-disentangled embedding sharing for further gains in sample efficiency (HE ET AL. 2021), the specific line Impacter Pathway's production DeBERTa-v3-base encoder sits on. CORAL, for its part, enforces rank-consistent, monotonic score bands by construction, through a weight-sharing constraint across the model's binary sub-classifiers, rather than leaving ordinal consistency to chance the way independent binary decisions would (CAO ET AL. 2020); it was introduced for a very different task, image-based age estimation, and has since become enough of a standard reference implementation for ordinal deep learning that it ships as an open-source library other teams reuse directly. In production, this pairing is Impacter Pathway's Model Suite v6.0, per the organization's published Model Card: a DeBERTa-v3-base encoder plus CORAL ordinal regression head, 184 million parameters, run under frozen, versioned model releases, with an average quadratic-weighted kappa of 0.918, average exact-match accuracy of 77.3%, and average adjacent accuracy of 98.0% (IMPACTER EDUCATION CORP. 2026). A given student response passes through the model once, deterministically, and produces one ordinal score; the architecture performs no sampling at inference and returns no second, different answer on a second pass over the same text. That is a categorically different failure surface than the one Ober et al. and Turpin et al. document, because it is not a generative model reasoning its way to

an answer. It is a fixed function evaluated once against a rubric-trained ordinal scale. Encoder-compression research points at the same design space from a different angle: DistilBERT, produced by applying knowledge distillation during pretraining rather than only at fine-tuning time, cuts parameter count by roughly 40 percent and inference time by roughly 60 percent while retaining about 97 percent of the larger model's language-understanding performance (SANH ET AL. 2019). That result matters here less for the compression itself than for what it demonstrates about the family this section is describing: encoder architectures in this lineage are well-studied enough that researchers can trade size for speed in predictable, measured increments, which is a different kind of engineering maturity than the reproducibility problems documented above in generative, decoder-based coding. The encoder-plus-ordinal-head choice is a deliberate point in that well-studied design space, not an ad hoc one.

The distinction that follows from all of this is worth stating plainly rather than leaving implicit. The clinical NLP literature above establishes what language-based detection of a diagnostic-adjacent construct like depression can currently achieve in the aggregate, across many studies, many populations, and many methods, a ceiling that, per Fisher et al., tops out around 80 percent accuracy even for the single best-studied construct in the field. Impacter Pathway's scoring pipeline is not making that claim, and is not trying to clear that particular bar, because it is not answering that question. It is scoring authentic, multilingual student language against Impacter Pathway's own rubric-defined ordinal bands for durable-skills constructs, using a deterministic, versioned model built specifically to avoid the reproducibility failure mode the construct-coding literature documents in generative alternatives. That is a narrower, more tightly specified measurement task than clinical detection: the construct is Impacter Pathway's own, rubric-defined and expert-authored rather than inherited from a diagnostic manual; the scale is ordinal rather than binary; and the architecture is discriminative rather than generative. Evaluated on its own terms, that narrower task is exactly what Section 3 turns to next.

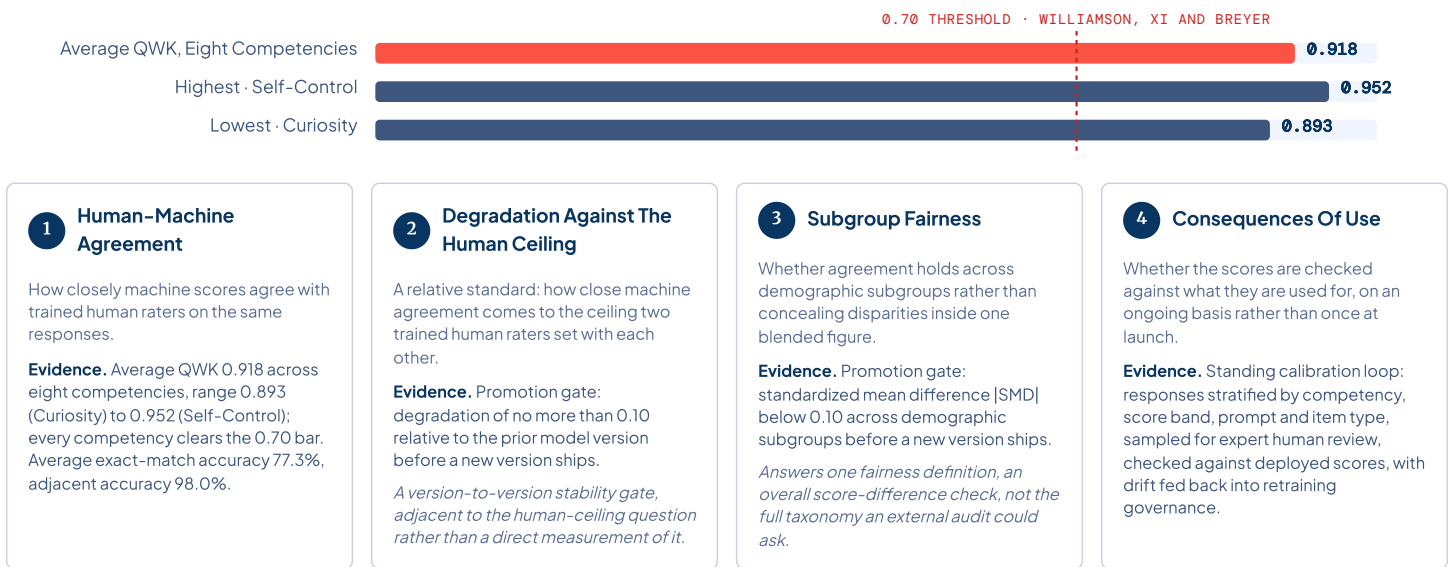
SECTION 03 · TRUSTWORTHY AUTOMATED SCORING

Trustworthy Automated Scoring: The Evaluation Standards

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.1: the published standards automated scoring is held to: agreement with trained human raters, degradation against human-human agreement, subgroup fairness, and the consequences of a scoring decision.

Human-Machine Agreement Against The Field's Threshold



WHAT THE SECTION COVERS

- The Field's Reference Framework
- Converging Standards, And The Umbrella Above Them
- The Lineage Behind An Agreement Statistic
- What Agreement And Fairness Don't Answer
- Impacter Pathway's Architecture And Figures Against The Standard

The full section, with every source beside the prose: bhpm.care/deliverables/d2/automated-scoring

BOTTOM LINE · THE RECORD'S OWN SUMMARY

Impacter Pathway's automated scoring holds up against the field's own evaluation standards, not against its own description of itself. Benchmarked to the Williamson/Xi/Breyer framework and its .70 QWK threshold, the Model Card figures clear the bar rather than merely assert it: an average QWK of 0.918, all eight competencies above 0.70, and promotion gates of QWK ≥ 0.70 , degradation ≤ 0.10 , and |SMD| < 0.10 before any new model version ships. The continuous stratified-sampling calibration loop keeps that bar met on an ongoing basis, not as a one-time validation claim frozen at launch. Every raw agreement number here carries the pedigree Cohen and Krippendorff's work established, not a decorative statistic standing alone. Read against the Loukina, Madhani, and Zechner fairness taxonomy, the |SMD| gate answers one specific, narrower fairness question: whether a group's overall score runs systematically higher or lower than another's. It does not answer the harder question the taxonomy also poses: whether a subgroup's score pattern still holds once true competency level is accounted for. That is a real limit on what the gate proves, not a claim that the gate settles the fairness question outright.

SECTION 03 · TRUSTWORTHY AUTOMATED SCORING · CONTINUED

Trustworthy Automated Scoring: The Evaluation Standards

Hand an outside evaluator a scoring pipeline and the request that comes back is never "trust us." It is a specific stack of documents. A technical validation report with human-machine agreement statistics, broken out by subgroup rather than blended into one flattering number. A bias or fairness audit showing whether a system trained on one population misreads the language of another. A construct definition or rubric the model was actually trained against, so the evaluator can check a score against the thing it claims to measure rather than against the model's own internal consistency. A version history explaining what changed between model releases and why, and whether performance moved in the process. And a description of how the system is monitored once it is in production, not just how it performed on a held-out test set the week it shipped. None of that is exotic to ask for. It is the same paper trail an evaluator would expect from a human-rater scoring operation, adapted to a machine rater: proof the rater was trained against a defined standard, proof its judgments were checked, proof its judgments were checked again once it went to work, and proof someone is still watching. That is not an idiosyncratic wish list a particular reviewer happens to favor. It is close to a checklist assembled from a specific, decades-old body of measurement literature that the educational testing field built precisely so evaluators would not have to invent their own criteria case by case, and any automated scoring system entering K-12 behavioral health reporting has to answer to that literature rather than to a vendor's description of its own model.

The Field's Reference Framework

The field's reference framework comes from [WILLIAMSON ET AL. 2012](#), which set out the evaluation criteria now treated as the baseline for any automated scoring of constructed-response tasks: the degree of agreement between machine and human scores, the size of any degradation relative to human-human agreement on the same responses, whether that agreement holds up across subgroups rather than concealing disparities inside an aggregate figure, and, the criterion most often skipped by vendors, an explicit accounting of the consequences that follow from a given score. The degradation criterion is easy to misread as a simple pass-fail number and is not one. It is a relative standard: the target is not that a machine clears some fixed bar in isolation, but that it comes close to matching how well two trained human raters agree with each other on the same responses, because that is the ceiling any scoring process, human or automated, realistically operates under. Williamson, Xi, and Breyer frame a quadratic-weighted kappa of roughly .70 as the industry benchmark for acceptable human-machine agreement, a threshold downstream automated-scoring work has treated as a floor rather than a target, but the framework's more demanding contribution is

the instruction to keep asking what a given score is used for and whether the evidence on hand justifies that use, not to stop once a single reliability number clears a bar.

Converging Standards, And The Umbrella Above Them

Two independent testing organizations have since converged on compatible guidance, from different institutional vantage points. [MCCAFFREY ET AL. 2022](#), ETS's institutional best-practices report for constructed-response scoring, folds automated and human-rater scoring into one shared quality framework rather than treating machine scoring as a separate, lesser-governed discipline. The report's premise is that scoring quality is a property of an entire system, rubric design, rater or model training, adjudication procedures, and ongoing quality control together, not a property of the algorithm alone, and it extends that guidance explicitly to multimodal responses rather than written text alone. [WOOD ET AL. 2021](#) arrives at the same place from the practitioner side, synthesizing the automated-scoring literature into a public, unpaywalled standards document that frames scoring quality as an ethical obligation of the professionals who build and deploy these systems, not merely a technical specification a vendor can satisfy quietly and move on. That it is freely posted, with no institutional login or licensing fee standing between it and a school board member trying to check a vendor's claims, is itself part of what makes it usable as public accountability evidence rather than an internal industry reference. Both documents also make a point specific enough to be worth stating on its own: nothing about a scoring process becomes exempt from these standards because a model, rather than a person, is applying the rubric. The same rater-training, adjudication, and quality-control logic that governs a room of trained human scorers governs a scoring model, and McCaffrey and colleagues write their guidance as a single document precisely so that a system cannot substitute a strong agreement statistic for the rest of that governance chain.

Sitting above both is the umbrella standard for the entire testing profession, [AMERICAN EDUCATIONAL RESEARCH ASSOCIATION ET AL. 2014](#), which anchors validity, reliability and precision, and fairness as the three pillars any testing practice, automated scoring included, has to address. None of the three is optional or substitutable for the other two: a system can post an impressive agreement statistic and still fail the Standards if the score does not mean what it claims to mean for the population it is used on, or if subgroups experience systematically different error patterns the aggregate number hides. The 2014 edition was explicitly revised to reckon with the growing role of technology in scoring decisions, which is precisely the developmental moment automated scoring for behavioral health reporting now sits inside. A pipeline used for public reporting has to answer to all three documents above and to this umbrella standard, not cherry-pick whichever framework happens to flatter a particular model's strongest metric.

SECTION 03 · TRUSTWORTHY AUTOMATED SCORING

The Lineage Behind An Agreement Statistic

Agreement is not a single number so much as a family of statistics with a specific history, and getting that history right matters because it is what makes a QWK figure interpretable rather than decorative. Reliability and agreement statistics used across this literature trace back to **COHEN 1960**, which introduced kappa as a chance-corrected measure of rater agreement, correcting for the fact that two raters will agree some of the time by chance alone even if neither is paying close attention. That is the mathematical basis the quadratic-weighted variant later extended for ordinal scales, and the weighting is not a cosmetic detail: a quadratic penalty means a machine score that misses a human rating by one rubric point costs the statistic far less than a miss of two or three points, which is exactly the property a competency rubric needs, since a one-point miss and a three-point miss are not equally serious errors for a district reading the result.

KRIPPENDORFF 2004 pushed the field further by cataloguing how badly percent-agreement and simple correlation are routinely misused as stand-ins for reliability, and by generalizing agreement measurement, via Krippendorff's alpha, across nominal, ordinal, and interval scales, more than two raters, and incomplete data. Those are conditions that describe most real scoring operations, including large-scale student response review, far better than the two-rater, complete-data case simpler statistics assume, and Krippendorff's own interpretive benchmarks, roughly .80 for a reliability figure worth relying on and .667 as the floor below which conclusions should be treated as merely tentative, are a stricter yardstick than a single favorable QWK number reported in isolation. Citing agreement without this lineage invites exactly the skepticism a reviewer should have: a QWK of .90 means something specific only because the statistic itself has a defensible pedigree and survives being checked against more than one reliability lens. This is also why an adjacent-accuracy figure, the share of scores that land within one rubric point of a human rater's judgment even when they do not match exactly, belongs in the same family of evidence rather than as a separate, softer statistic reported only when the exact-match number looks less flattering on its own. Read through Cohen and Krippendorff's lineage, adjacent accuracy is simply another way of asking the same weighted-agreement question the quadratic-weighted kappa asks, restated as a plain percentage a non-statistician reviewer can check by hand.

What Agreement And Fairness Don't Answer

Fairness deserves the same scrutiny agreement gets, and the literature treats it as considerably less settled. **LOUKINA ET AL. 2019**, from ETS researchers working on automated scoring for a large-scale English-proficiency assessment, lays out several distinct and only partly compatible definitions of fairness for a scoring system, using a test-taker's native-language background as the demographic variable under study across both simulated and real data. One version of fairness asks whether the system is roughly as accurate for every group on average. A second asks whether the system's scores run systematically higher or lower for one group than another. A third, more demanding version asks whether that group-level difference survives once the groups' actual underlying proficiency is accounted for, rather than being an artifact of one group genuinely performing differently on the construct being measured. The paper's central and

somewhat uncomfortable finding is that total fairness, in the sense of satisfying every one of these definitions at once, may not be achievable: a change that closes one gap can open another, and different stakeholders reasonably prioritize different definitions depending on what the score is used for. That matters for reading Impacter Pathway's own promotion-gate language, which requires a standardized mean difference of less than 0.10 across the demographic subgroups the fairness check screens before a new model version can ship. That gate operationalizes something close to the second, simpler fairness definition in Loukina and colleagues' taxonomy, an overall-score-difference check. It does not, by itself, answer the third and harder question: whether a subgroup's score pattern still looks fair once true competency level is held constant, rather than only checked as a raw group average. That is a legitimate, more demanding question for an external evaluator to ask of any automated scoring system, Impacter Pathway's included, and it is exactly the kind of question the document request an evaluator opens with is built to surface rather than something a single passed threshold closes off.

Agreement and fairness statistics answer whether a system's numbers behave well. They do not by themselves answer whether the system is measuring the thing it claims to measure, which is where **BAUER ET AL. 2020** supplies the argument that separates a defensible scoring system from a merely well-fitted one. Bauer and Zapata-Rivera argue that automated scoring has to be traceable to an explicit cognitive model of the construct an assessment is meant to elicit, not just a statistical fit between machine output and human ratings, and that the transparency of a scoring system's internal logic is itself a validity concern, not a usability nicety layered on afterward. A model can post excellent agreement statistics against human raters who were themselves scoring from a vague or contested rubric, and the resulting number would still not mean what a district needs it to mean. That is the same claim Ober and colleagues make from the language-model side in the discussion of construct clarity elsewhere in this review: what makes scoring trustworthy is that a rater, human or machine, is applying a legible rubric to a legible construct, not that the underlying model is larger, newer, or more complex. It is the standard Impacter Pathway's own architecture is designed to meet directly rather than approximate.

Impacter Pathway's Architecture And Figures Against The Standard

Impacter Pathway's Model Suite v6.0 scores each of the eight durable-skills competencies with a fine-tuned DeBERTa-v3-base encoder, 184 million parameters, paired with a CORAL ordinal-regression head. The encoder descends from the ELECTRA-style pretraining and gradient-disentangled embedding-sharing architecture described in **HE ET AL. 2021**, which replaced the masked-language-model pretraining used by earlier BERT-family models with a more sample-efficient replaced-token-detection task and resolved a training instability, described in the paper as a tug-of-war between generator and discriminator objectives, that had limited how far that more efficient pretraining approach could be pushed.

SECTION 03 · TRUSTWORTHY AUTOMATED SCORING · CONTINUED

Impacter Pathway's Architecture And Figures Against The Standard

Neither change is cosmetic: they are what let a base-sized, 184-million-parameter encoder reach the level of language understanding a scoring pipeline needs without the cost and latency of a much larger model. That sizing choice is itself worth naming as a choice rather than a limitation. A base-sized encoder is smaller than the large or XL DeBERTa-v3 variants the same paper reports, and choosing it trades some raw language-modeling headroom for a model that is cheaper to run at the volume a statewide behavioral health reporting pipeline requires and easier to audit end to end, a tradeoff that only makes sense once a construct is well-specified enough, in Bauer and Zapata-Rivera's sense, that the model does not need to be enormous to represent it well. The CORAL framework, from [CAO ET AL. 2020](#), matters here for a specific reason: it was built to guarantee rank-monotonic, internally consistent predictions across ordinal categories, which is the exact property a competency-level rubric score needs and a generic classification head does not provide. It is worth noting that CORAL was not developed for language scoring at all; its originating application was estimating a person's age from a photograph, an unrelated ordinal-prediction problem. That Impacter Pathway's architecture borrows a general, independently peer-reviewed ordinal-regression method built for a different domain, rather than a bespoke, unaudited scoring head built in-house, is itself a piece of the evidence an evaluator's document request is checking for: a documented, externally verifiable methodological choice rather than an internal claim taken on faith.

Per the Impacter Pathway Model Card [IMPACTER EDUCATION CORP. 2026](#), this architecture produces an average quadratic-weighted kappa of 0.918 across the eight competencies, an average exact-match accuracy of 77.3%, and an average adjacent, within one score point, accuracy of 98.0%. Per-competency QWK ranges from 0.893 on Curiosity to 0.952 on Self-Control, with every one of the eight competencies clearing the 0.70 threshold Williamson, Xi, and Breyer's framework treats as the field's minimum acceptable bar, most of them by a wide margin, and the spread between the lowest- and highest-performing competency is itself a piece of subgroup-style evidence worth stating plainly rather than folding into a single

blended average. New model versions are held to explicit promotion gates before they ever reach production: quadratic-weighted kappa at or above 0.70, degradation of no more than 0.10 relative to the prior version, and the standardized mean difference of less than 0.10 across demographic subgroups discussed above. A version that fails any gate does not ship. Once a version clears the gates, it is frozen and versioned for deterministic inference, so the same student response scored twice against the same model version returns the same score, an auditability property a generative, sampling-based system cannot make the same guarantee about, and one that matters directly for the reproducibility concerns documented elsewhere in this review's discussion of generative construct coding.

Meeting a gate once is not the same as staying trustworthy over time, which is why evaluation under this framework has to be continuous rather than a one-time certification exercise, the same point Williamson, Xi, and Breyer make about consequences generally and the same gap McCaffrey and colleagues' shared quality framework is built to close for ongoing operations rather than launch-day validation alone. Impacter Pathway's production calibration architecture, documented internally as the Model Suite v6.0 technical note [IMPACTER EDUCATION CORP. 2026](#), implements this as a standing operational loop rather than an annual audit: student responses are stratified by competency, score band, prompt, and item type, a sample is drawn from each stratum for expert human review, that review is compared against the deployed model's scores, and any detected drift feeds back into retraining governance before it reaches a threshold that would require a new promotion cycle. That loop is the mechanism that keeps the Model Card's figures current rather than a snapshot that quietly goes stale as new prompts, cohorts, and item types enter production, and it is the kind of ongoing-monitoring documentation the evaluator in the opening scenario is asking for when the request is for more than a single validation report dated the week the model shipped.

Laid out side by side, the Williamson/Xi/Breyer framework's four criteria and the evidence already cited above for each map onto a single checklist:

Evaluator demands

What an independent evaluator asks of a process-evaluation record. It is drawn as a diagram in the earlier working design.

[Open evaluator demands in the archive](#)

SECTION 04 · PROCESS & OUTCOME INDICATORS

Process And Outcome Indicators In MTSS-Aligned Behavioral Health

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.1 with Activity 2.3: the candidate indicator set, split into process and outcome families and sorted into the Commission's four reporting domains: services provided; student demographics and characteristics; service utilization; outcomes and impact.

Thirteen Candidate Indicators, By Family And Commission Domain

From the section's indicator table. Process and outcome are separate, independently reportable families; contextual and utilization indicators sit between them.

1 Services Provided4	2 Students Served3	3 Service Utilization3	4 Outcomes And Impact3
PROCESS Screening Administration Completion Rate	CONTEXTUAL Screened Population Profile (Grade Band, Language, Program Status)	UTILIZATION Referral-To-Service Rate	OUTCOME Tier Movement Over Time
PROCESS Implementation Fidelity Score (TFI-Aligned)	CONTEXTUAL Baseline Tier And Need-Level Distribution	UTILIZATION Billable Screening Encounters (CPT 96127)	OUTCOME COVARIATE Exclusionary Discipline Rate (Suspensions And Expulsions)
PROCESS Staff And Educator Training Participation Rate	CONTEXTUAL English Learner And IEP/504 Status Cross-Tab On Screening Participation	UTILIZATION School-Linked Provider Network Utilization	OUTCOME Domain-Level Growth On Measured Competencies
PROCESS Data-Review Cycle Completion			

WHAT THE SECTION COVERS

- School-Wide Positive Behavior Support Is A Systems-Level Practice, Not One Procedure
- A Validated Instrument For Scoring Fidelity
- The Outcome Side Of The Ledger: What The One Randomized Trial Found
- Three Indicator Families, Not One Blended Score
- Service Utilization: Precedent From Two Comparable Initiatives

The full section, with every source beside the prose: [bhpmmcare/deliverables/d2/indicator-families](#)

BOTTOM LINE · THE RECORD'S OWN SUMMARY

Process and outcome are two separate, independently reportable indicator families, not one blended score. The Scott-and-Gage fidelity study, the Horner-Sugai-Anderson evidence-quality review, and the MTSS-B randomized trial converge on that distinction, even though they disagree on how strong any single fidelity-outcome association actually is.

The indicator set above is the working skeleton, already mapped to all four Commission domains. Service-utilization indicators carry a feasibility flag rather than a promise, because each rests on a site's CYBHI enrollment and billing infrastructure.

Because the MTSS-B trial's real effects were concentrated in a highest-need subgroup, disaggregation by subgroup and need tier is a priority for the outcomes domain, not an optional add-on.

The Project AWARE and CYBHI Fee Schedule figures are comparable reporting precedent for what the service-utilization domain can look like, not evidence that BHPM itself has produced those results.

This literature informs the indicator set's structure and weighting; it does not by itself finalize which indicators ship.

SECTION 04 · PROCESS & OUTCOME INDICATORS · CONTINUED

Process And Outcome Indicators In MTSS-Aligned Behavioral Health

A four-year study tracking twenty-nine schools implementing Multi-Tiered Systems of Support found something specific enough to build a measurement architecture on: schools with higher fidelity in the behavior-support domain of their MTSS implementation logged significantly fewer suspension events than matched comparison schools, while schools with higher fidelity in the academic domain, reading and math instruction delivered through the same tiered structure, showed no such reliable relationship to achievement gains (SCOTT ET AL. 2019). Scott, Gage, Hirn, Lingo, and Burt reached that finding using the Academic and Behavior Response to Intervention School Assessment, an instrument built to score fidelity separately on behavior-domain and academic-domain subscales rather than producing one blended implementation score, and they tracked those subscales against suspension counts and state achievement data across the same twenty-nine schools each year for four years. Their own conclusion is blunt: school-wide strategies that do not specifically involve effective instruction in academic areas are unlikely to raise academic achievement on their own, no matter how faithfully the behavior-support scaffolding around that instruction is run. That is not a caveat buried in a footnote. It is the finding that decides how BHPM's indicator set has to be built: fidelity and outcome are not the same measurement, one domain's fidelity does not transfer its predictive power to a different outcome domain, and an indicator system that collapsed "the program ran as designed" into "the program worked" would have been wrong about exactly the distinction this study was built to test.

School-Wide Positive Behavior Support Is A Systems-Level Practice, Not One Procedure

That separation is easier to state than to operationalize, because school-wide positive behavior support is not a single discrete practice with a clean efficacy-trial literature behind it. It is a systems-level, multi-component framework layering universal screening, tiered response, data-based decision rules, and ongoing coaching across three tiers of intensity. Horner, Sugai, and Anderson confronted that complexity directly by proposing six criteria for defining any educational practice clearly enough to assess its evidence base at all: that the practice be operationally defined, that the settings where it is expected to work be specified, that its target population be defined, that the qualifications required to use it successfully be defined, that its expected outcomes be defined, and that the conceptual theory explaining why it should work be stated (HORNER ET AL. 2010). Applied to SWPBS, those criteria forced the reviewers to treat the practice as a constellation, three tiers of intervention plus the organizational systems that sustain them, rather than pretend it reduces to one procedure with one efficacy trial behind it. Their evidence review covered 46 peer-reviewed studies published between 2000 and 2009, split across the three tiers, twenty at Tier 1, thirteen each at Tiers 2 and 3, and graded against a

second, five-part standard: whether the practice and participants were defined with enough precision to replicate, whether the measures used were valid and reliable, whether the designs were rigorous, whether documented effects came without harmful side effects, and whether those effects held up over time. Their conclusion, tier by tier, is where the caution BHPM needs actually lives. Tier 1 implementation fidelity was found to be achievable at meaningful scale, elementary school teams receiving four to six days of training reached eighty percent or better fidelity on a standardized evaluation tool in a randomized waitlist-controlled trial, and associated with real reductions in problem behavior and improved organizational health. On academic outcomes specifically, the review's own language is deliberately hedged: SWPBS implementation is "promisingly, but not definitively" associated with increased academic outcomes, and the authors state outright that it is premature to claim SWPBS is causally associated with improved academic achievement, because the underlying mechanism runs through improved instructional engagement rather than through behavior support directly producing learning. That caution, published independently of Scott and Gage's later fidelity study, arrives at close to the same place from a different kind of evidence, a synthesis of dozens of studies rather than one four-year panel, converging on the conclusion that behavior-domain fidelity and academic outcome are not the same claim and should not be reported as though they were.

The same review flags two further limitations worth carrying into BHPM's own feasibility notes rather than smoothing over. Tier 2 and Tier 3 evidence rests overwhelmingly on single-case designs rather than the randomized and quasi-experimental group trials available for Tier 1, a real difference in methodological weight that a performance system reporting all three tiers on one dashboard should not obscure. And implementation at meaningful fidelity is measurably harder in high schools than in the elementary settings most of this evidence base was built on, a pattern the reviewers describe as consistent across the roughly 1,000 high schools that had adopted SWPBS by the time of their review. Neither limitation is a reason to abandon a tiered indicator design; both are reasons a site-level fidelity report should say which tier and which school level it is describing rather than publishing one undifferentiated score.

A Validated Instrument For Scoring Fidelity

Fidelity itself needs a validated instrument before any of this can be scored, and the field-standard measure is the Tiered Fidelity Inventory, developed and maintained under the OSEP-funded Center on PBIS specifically to assess fidelity across Tier 1 through Tier 3 in one linked tool rather than three unrelated instruments. McIntosh and colleagues established its technical adequacy across three separate linked studies: a content-validity study confirming the instrument's items actually represent the practices each tier is supposed to measure, a usability and reliability study, and a large-scale validation drawing on a broad multi-state sample

(MCINTOSH ET AL. 2017).

SECTION 04 · PROCESS & OUTCOME INDICATORS · CONTINUED

A Validated Instrument For Scoring Fidelity

Across those three studies the TFI showed strong construct validity at all three tiers, strong interrater agreement, stable two-week test-retest reliability, and strong correspondence with the SWPBIS fidelity measures that predated it. That last property matters for a system trying to connect to existing district practice: a site that already has fidelity data collected under an older SWPBIS measure is not starting from zero when it adopts TFI-aligned scoring, because the two are validated against each other rather than measuring unrelated constructs.

The Outcome Side Of The Ledger: What The One Randomized Trial Found

The outcome side of the ledger carries its own, sharper caution, drawn from the one independent randomized controlled trial in this space large enough to trust. An IES-funded evaluation of MTSS-B, the behavior-focused variant of Multi-Tiered Systems of Support, randomized roughly eighty-nine elementary schools across nine districts that had not previously implemented PBIS, and its results are not the clean success story a vendor slide would prefer

MCINTOSH ET AL. 2023. Averaged across all students, the program did not improve disruptive behavior, other behavior outcomes, or achievement. Where effects appeared, they were concentrated entirely among the roughly fifteen percent of students identified at baseline as struggling most with behavior: that group showed improved reading scores and reduced disruptive behavior, but only by the second year of full implementation, not the first. Some broader improvements in classroom management and school climate measures did appear at the school level. And even the one clear win did not fully hold: the reading effect for the highest-need students did not persist into the follow-up year after program support ended. An indicator system built only to detect aggregate, school-wide movement would miss this trial's entire finding, because the finding lives inside a subgroup that a school-wide average erases, and it would also miss the timing: an outcome measured only in year one of implementation would have reported a null result the trial's own second-year data contradicts. BHPM's outcome and contextual indicators are therefore designed to be disaggregated by subgroup and need tier from the start, not retrofitted for equity reporting after the fact, and to be reported on a timeline long enough to distinguish a first-year null from a second-year effect rather than closing the book early.

Three Indicator Families, Not One Blended Score

Read against each other, these three sources do not fully agree, and the disagreement is itself informative. Scott and Gage's finding comes from a quasi-experimental panel comparing schools against matched comparisons, not random assignment; Horner, Sugai, and Anderson's caution about academic outcomes is a synthesis judgment drawn across dozens of underlying studies of uneven design quality; and the MTSS-B brief is the one true randomized trial

in the group, the design least vulnerable to the selection effects that can inflate fidelity-outcome associations in a matched-comparison study. That the RCT's own aggregate result is null while the quasi-experimental and synthesized evidence report real associations is not a contradiction to paper over. It is exactly the pattern a well-built indicator system should expect and survive. A system that reported only the matched-comparison finding would overstate what fidelity data alone can promise; a system that reported only the RCT's null aggregate result would erase the subgroup effect that same trial documents. BHPM's three-family indicator structure (process measured on its own terms, outcome measured and disaggregated on its own terms) is a direct response to that pattern, not an attempt to resolve it into one number.

Three indicator families follow from that literature, and the boundaries between them track the design tension above rather than arbitrary bookkeeping. **Process indicators** measure whether the system is running as designed: screening completion, implementation fidelity scored against a validated instrument, training participation, and data-review cadence, the domain where Scott and Gage's finding and the Horner-Sugai-Anderson framework criteria are strongest. **Outcome indicators** measure what the system is meant to produce: tier distribution and movement over time, domain-level growth on measured competencies, and covariates like attendance and exclusionary discipline that the fidelity and RCT literature both link, in different ways, to behavior-domain implementation. **Contextual indicators** describe who is being measured and who is receiving services: demographic composition of the screened population, baseline tier and need-level distribution, and service-receipt cross-tabs by subgroup, the domain the MTSS-B trial's subgroup-concentrated effect makes non-optional rather than a nice-to-have. None of the three families is optional. A process-only report cannot answer whether the work mattered; an outcome-only report, especially one that only ever reports the school-wide average, cannot answer for whom it mattered or explain why results moved.

Those three families map directly onto the four reporting domains the Commission has asked for: services provided, student demographics and characteristics, service utilization, and outcomes and impact. Process indicators populate the services-provided domain. Contextual indicators populate the demographics domain and travel alongside outcome indicators wherever disaggregation is required, the direct design response to the MTSS-B trial's subgroup finding above. Service utilization sits partly inside BHPM's own platform data and partly inside the state and federal reporting infrastructure already built around school-based behavioral health billing, which is why two comparable initiative reporting frames, not academic studies, anchor that domain below rather than a fifth research citation. Outcome indicators close the loop back to the Commission's central accountability question, disaggregated on the timeline the RCT evidence above argues for rather than collapsed into a single annual figure.

SECTION 04 · PROCESS & OUTCOME INDICATORS · CONTINUED

Three Indicator Families, Not One Blended Score

The diagram below re-sorts the same working indicator set by the Commission's own four domain names and order, rather than by the three literature-driven families above. In this candidate set the two groupings line up one-to-one (process indicators are the services-provided domain, contextual indicators are the demographics domain, and the outcome and outcome-covariate indicators

together are the outcomes domain), but the families and the domains answer different questions and won't necessarily stay aligned as the indicator set grows: the families explain why an indicator belongs in the literature base, the domains are how the Commission has asked to receive it.

Domain alignment

The candidate indicator set sorted into the Commission's four reporting domains. It is drawn as a diagram in the earlier working design.

[Open domain alignment in the archive](#)

The table below is the working candidate indicator set, and it is the skeleton BHPM's indicator reporting is built on rather than an illustrative sample: each row states the literature basis for including the indicator, the collection specification a site would actually

implement, the Commission domain it reports into, and a feasibility note distinguishing what BHPM can measure today from what depends on a data-sharing agreement or a district's own billing infrastructure. Clicking a row opens the full profile.

INDICATOR	INDICATOR FAMILY	COMMISSION DOMAIN	LITERATURE BASIS	COLLECTION SPEC	FEASIBILITY NOTE
Screening administration completion rate	Process	Services provided	Fidelity-to-outcome logic: implementation fidelity is a distinct, prior condition to outcome movement, not a proxy for it.	Platform-logged screening completions divided by enrolled eligible students, per school, per reporting window.	High: captured automatically by the platform; no manual entry.
Implementation fidelity score (TFI-aligned)	Process	Services provided	Tiered Fidelity Inventory is the field-standard, psychometrically validated tiered fidelity-of-implementation measure.	Periodic TFI or TFI-aligned local fidelity check scored across Tier 1–3 domains by a trained site coordinator.	Moderate: requires a trained local scorer and a fixed administration cadence.
Staff and educator training participation rate	Process	Services provided	Reporting category drawn from a comparable federal school behavioral health program's own public accounting of training reach.	Count of staff completing BHPM onboarding and technical-assistance modules divided by the site staff roster.	High: logged through existing TTA delivery tracking.
Data-review cycle completion	Process	Services provided	Fidelity measurement frameworks for tiered systems treat a regular data-review routine as a scoreable implementation component, not an assumed byproduct of adoption.	Count of scheduled site data-review meetings held, against the cadence set in the site's implementation plan.	High once a site's review calendar is established in DI.
Screened population profile (grade band, language, program status)	Contextual	Student demographics and characteristics	Independent RCT evidence that effects concentrate among highest-need students argues for reporting the screened denominator's composition, not just its size.	Enrollment-derived demographic fields cross-tabulated against screening completion, aggregated to school level.	High: fields already reside in roster feeds.
Baseline tier and need-level distribution	Contextual	Student demographics and characteristics	Tiered-system framing for evaluating school-wide behavior support treats tier placement as the organizing unit of analysis.	Count of students at each MTSS tier at first screening administration, by school.	High.
English learner and IEP/504 status cross-tab on screening participation	Contextual	Student demographics and characteristics	Subgroup-concentrated effect sizes in the IES MTSS-B trial motivate disaggregating participation, not only outcomes, by subgroup.	Screening participation rate disaggregated by English learner and IEP/504 status.	Moderate: depends on the site's data-sharing agreement covering those fields.
Referral-to-service rate	Utilization	Service utilization	Comparable federal program reporting category: referral volume as a standing public accountability figure for a school behavioral health initiative.	Count of students referred for behavioral health service following screening, tracked to service entry where data-sharing allows.	Moderate: depends on district-provider data-sharing agreements.
Billable screening encounters (CPT 96127)	Utilization	Service utilization	CPT 96127 is the AMA-maintained billing code for a brief, standardized-instrument emotional/behavioral assessment with scoring and documentation.	Count of standardized-instrument screening encounters coded and documented under CPT 96127.	High where a site's billing infrastructure already supports the code; low where none exists yet.

SECTION 04 · PROCESS & OUTCOME INDICATORS · CONTINUED

Three Indicator Families, Not One Blended Score

INDICATOR	INDICATOR FAMILY	COMMISSION DOMAIN	LITERATURE BASIS	COLLECTION SPEC	FEASIBILITY NOTE
School-linked provider network utilization	Utilization	Service utilization	California's statewide CYBHI Fee Schedule Program, authorized under Welfare and Institutions Code Section 5961.4, is the in-state precedent for reporting reimbursed school-linked encounters.	Count and rate of reimbursed encounters under the DHCS fee schedule at CYBHI-enrolled sites.	Depends on the district's CYBHI cohort enrollment status.
Tier movement over time	Outcome	Outcomes and impact	Behavior-domain implementation fidelity is empirically associated with downstream movement in student-level indicators across a multi-year, multi-school sample.	Change in tier placement between baseline and each subsequent screening window, aggregated to school level with small-cell suppression.	High once two screening administrations exist.
Exclusionary discipline rate (suspensions and expulsions)	Outcome covariate	Outcomes and impact	Behavior-fidelity gains associate with reduced exclusionary discipline; an independent randomized trial separately found reduced disruptive behavior concentrated among the highest-need students.	District-reported suspension and expulsion counts, cross-referenced against BHPM tier and fidelity data.	High: already collected for state reporting.
Domain-level growth on measured competencies	Outcome	Outcomes and impact	Applies the same process/outcome separation established in the MTSS fidelity literature to Impacter Pathway's own longitudinal attribute scores.	Within-student change in scored attribute levels across administrations.	High: computed automatically by the scoring pipeline.
Thirteen candidate indicators, each mapped to an indicator family and one of the Commission's four reporting domains.					

Service Utilization: Precedent From Two Comparable Initiatives

Service utilization is the domain where BHPM's own platform data runs out fastest, because a referral or a billed encounter is only visible to a school-based measurement system when the receiving service sits inside the same reporting relationship. Two comparable initiative reporting frames were chosen precisely because they already report at this level in public, not because either one evaluates a program's causal effect. SAMHSA's Project AWARE, a federal grant program that builds school-based behavioral health infrastructure through state, county, and district partnerships rather than delivering services directly, reports 2018–2024 aggregate outcomes that read like a preview of what BHPM's own service-utilization and outcomes domains will eventually publish: 3,492 organizations entered formal partnership agreements, 1,509 state or local policy changes were attributed to the program, 1,384,460 school staff, families, and community members participated in related trainings, and 327,332 students were referred for mental health or related services [\(SUBSTANCE ABUSE AND MENTAL HEALTH SERVICES ADMINISTRATION. 2026\)](#). Those figures describe a federal grantee's own program accounting, the same category of reporting a comparable BHPM rollup would eventually produce, not a peer-reviewed evaluation of what the funding caused; SAMHSA's own page presents them as cumulative reach, not as an effect size against a comparison condition. BHPM's indicator set borrows only the reporting categories, partnerships formed, staff trained, students referred, not the causal claim, and the distinction is worth stating plainly because a partnership-agreement count and a randomized trial's effect size answer different questions even when they appear on the same kind of program dashboard.

California offers a state-level precedent with more legal specificity than a federal grant program's aggregate reach. The Children and Youth Behavioral Health Initiative's Fee Schedule Program, authorized under Welfare and Institutions Code Section 5961.4, requires commercial health plans and Medi-Cal to reimburse school-linked

behavioral health providers at published rates regardless of network status, effective January 1, 2024 [\(CAL. WELF. & INST. CODE § 5961.4 ET AL. 2024\)](#). DHCS has rolled the program out in cohorts: 484 local education agencies and five institutions of higher education across Cohorts 1 through 4, with Cohort 6 participants announced February 27, 2026 following an operational-readiness review covering each applicant's Medi-Cal enrollment status, service-delivery capacity, data-collection practices, and billing infrastructure [\(CALIFORNIA DEPARTMENT OF HEALTH CARE SERVICES. 2026\)](#). That readiness review is itself a useful precedent for BHPM's own feasibility notes: DHCS does not treat "the district joined the program" as equivalent to "the district can bill against it," and a BHPM service-utilization indicator should not either. The billing-code detail that ties screening activity to a documented, reimbursable service touchpoint is CPT 96127, the standardized code for a brief emotional or behavioral assessment with scoring and documentation, billed per standardized instrument administered, up to four units per encounter rather than by time spent, and commonly billed alongside instruments like the PHQ-9 and GAD-7 [\(AMERICAN MEDICAL ASSOCIATION. 2026\)](#). Wherever a BHPM site's own billing infrastructure already supports that code, screening volume and billed-encounter volume become the same countable event, the strongest possible link between a process indicator and a utilization indicator, and one that will only exist at sites whose CYBHI enrollment and billing systems have cleared the same operational bar DHCS itself applies.

None of this is a wish list dressed up as a table. Every row in the indicator set above states what BHPM can measure today under current site infrastructure and what depends on a data-sharing agreement, a district's CYBHI enrollment status, or the billing infrastructure a site runs. That distinction follows directly from the pattern the literature above establishes: fidelity data, synthesis judgments, and randomized evidence do not always agree with each other, subgroup effects can hide inside school-wide averages, and a performance management system that overpromises data it cannot produce, or certainty a single study cannot support, fails the same accountability test it exists to serve.

SECTION 05 · IMPLEMENTATION SCIENCE

Implementation Science And Process Evaluation

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.1: the implementation-science frameworks the program instantiates, from readiness assessment through the collaborative learning groups, and what an independent evaluator asks of that record.

How BHPM's Own Sequence Maps Onto The Staged Frameworks

1	2	3	4
Exploration	Installation	Initial Implementation	Full Implementation
Assess whether the innovation fits an identified need and whether the resources exist to support it, before anyone commits to it.	Put structural supports in place before first real use: staffing, funding lines, data systems, policy alignment.	The middle stage, where a new practice meets an old system and needs intensive support to survive first contact.	The new practice is integrated into a site's normal, ongoing routines, not a special project running alongside business as usual.
DELIVERABLE 1 1 2 Site Readiness Assessment Checked against CFIR's inner-setting domain as well as the driver framework, so the instrument asks about organizational context, not only technical capacity, before deployment.	DELIVERABLE 2 2 Configured Deployment The indicators, index and scoring infrastructure this brief documents: the decision-support data system Fixsen names as its own organization driver, not an add-on to service delivery.	DELIVERABLE 3 3 Training And Technical Assistance Supplies the ongoing coaching component the synthesis found training alone cannot substitute for.	DELIVERABLES 4 AND 5 4 Collaborative Learning Groups The networked-improvement-community mechanism Bryk and colleagues specify: a quarterly PDSA cadence with the composite index as a candidate common measure.

WHAT THE SECTION COVERS

- Implementation Is A Staged Process, Not An Event
- What Counts As Evidence That Implementation Succeeded
- Continuous Improvement Over Episodic Evaluation
- What The Evidence Says Works, Specifically, In School Mental-Health Implementation
- How BHPM's Own Sequence Maps Onto These Frameworks

The full section, with every source beside the prose: bhp.mcare.org/deliverables/d2/implementation-science

KEY TAKEAWAY · THE RECORD'S OWN SUMMARY

Process and fidelity indicators are organized around Proctor et al.'s eight-domain taxonomy and kept structurally separate from outcome indicators rather than blended into one scale, the same two-tier separation this literature draws throughout. The process-indicator family, completion, fidelity, training participation, data-review cadence, is the Fixsen/CFIR core-driver list: those are the specific levers this literature identifies as differentiating full implementation from stalled rollout, weighed against what the platform can actually observe. The quarterly Collaborative Learning Group cycle (D4/D5) is a PDSA-style networked improvement community in the Bryk et al. sense, with the composite index serving as the shared cross-site measure that carries what one quarter's cycle learns into the next. Because Baffsky et al.'s review finds its strongest evidence specifically for continuous monitoring joined to performance feedback on fidelity, not for periodic recall surveys and not, yet, as strongly for adoption, BHPM's platform analytics is the primary implementation-measurement instrument, with formal review points layered on top of that signal rather than substituting for it. That is the indicator architecture this literature supports.

SECTION 05 • IMPLEMENTATION SCIENCE • CONTINUED

Implementation Science And Process Evaluation

Picture a district that does everything a rollout checklist would tell it to do. It buys a validated screener, schedules a single August training day for every counselor across twelve schools, and sends staff back to their buildings with binders and good intentions. Eight months later, three schools are using the tool exactly as designed, five have quietly shortened it into something nobody wrote down, and four have gone back to the paper form they used the year before. Nothing about the screener changed. What changed is that training landed once and was never followed by coaching, and nobody was watching the data closely enough to notice the drift until an outside monitoring visit surfaced it months later. That pattern, strong content, weak delivery system, is close to the modal failure mode implementation science exists to explain, and the National Implementation Research Network's foundational synthesis names the underlying question directly: what actually separates programs that take hold in real service settings from those that stay confined to a manual (FIXSEN ET AL. 2005). This section lays out the frameworks BHPM draws on to make deployment itself measurable, not just the program being deployed, and argues that the project's own activity sequence, readiness assessment, configured deployment, technical assistance, collaborative learning, is a direct application of that literature.

Implementation Is A Staged Process, Not An Event

The Fixsen synthesis proposes that implementation unfolds across four identifiable stages rather than arriving in one step once a program is adopted. Exploration is the stage where a site assesses whether an innovation actually fits an identified need and whether the resources exist to support it, before anyone commits to it. Installation follows: the structural supports, staffing, funding lines, data systems, policy alignment, get put in place before the first real use, on the theory that a site should not be expected to build its own plane while flying it. Initial implementation is the awkward middle stage, where a new practice meets an old system and needs intensive support to survive first contact with real caseloads and real schedules. Full implementation is reached only when the new practice becomes integrated into a site's normal, ongoing routines, not a special project running alongside business as usual but simply how the work gets done. Programs can and do stall at any of the first three stages indefinitely; the synthesis treats that as the expected risk to manage against, not an edge case.

Within that staged model, the synthesis groups the specific practices that predict whether a site actually reaches full implementation into three interacting categories of "implementation drivers." Competency drivers, staff selection, training, coaching, and performance evaluation, build the capacity of the people doing the work. Organization drivers, systems intervention, facilitative administration, and a decision-support data system, build the capacity of the organization around them. Leadership drivers, technical and adaptive, hold the other two together when conditions get complicated. Two findings matter directly for a multi-district

behavioral health system. First, training without ongoing coaching and a data-feedback loop rarely produces full implementation on its own; a single August in-service is a competency-driver input with no organization-driver support behind it, close to a textbook description of why the district above stalled. Second, a decision-support data system is a core organization driver in its own right, not an add-on layered onto the "real" work of service delivery. A performance management platform generating usage and fidelity data as a byproduct of normal operation is, in this framework's own terms, implementation infrastructure rather than a separate reporting obligation.

A complementary, more recent framework fills a gap the 2005 synthesis leaves underspecified: why the identical program, delivered with identical training, succeeds at one site and stalls at another. The Consolidated Framework for Implementation Research organizes the contextual determinants of implementation success into five domains, intervention characteristics, outer setting, inner setting, individual characteristics, and implementation process, consolidating constructs from nineteen earlier implementation theories into one shared taxonomy rather than a new causal model (DAMSCHRODER ET AL. 2009). Its authors describe it as operationalizing the multi-level influence the Fixsen synthesis already gestures at, giving that framework's driver categories more granular constructs to sit under. For a project spanning six LEAs with different staffing, prior behavioral-health infrastructure, and district leadership, the inner-setting domain, implementation climate, leadership engagement, available resources, is close to what Deliverable 1's readiness assessment is built to capture before installation. Naming it explicitly is a useful check on that instrument: it should ask about organizational context, not only technical capacity, since CFIR treats the two as separable predictors that do not move together.

What Counts As Evidence That Implementation Succeeded

A parallel line of work addresses a definitional problem that follows directly from the staged model: if implementation is its own process, what does it mean to say that process succeeded, independent of whether the practice ultimately changed outcomes for students? Proctor and colleagues answer by proposing eight conceptually distinct implementation outcomes, defined precisely enough to be measured rather than gestured at. Acceptability is the perception among stakeholders that a practice is agreeable or satisfactory. Adoption is the intention or decision to try a practice in the first place. Appropriateness is its perceived fit for a given setting or population. Feasibility is whether it can be carried out within an agency's real constraints. Fidelity is the degree to which it was delivered as the developer intended. Implementation cost is the resource impact of the effort itself, separate from the underlying service's cost. Penetration is how deeply a practice has spread into a service setting and its subsystems, not just whether it exists somewhere in the building. Sustainability is whether it survives past its initial funding or novelty period and becomes part of ongoing operations

(PROCTOR ET AL. 2011).

SECTION 05 · IMPLEMENTATION SCIENCE · CONTINUED

What Counts As Evidence That Implementation Succeeded

Those eight sit inside an explicit three-tier logic, and the tiers matter as much as the definitions. Implementation outcomes are necessary but not sufficient evidence of service-system outcomes, which the paper anchors to the Institute of Medicine's quality dimensions, care that is safe, effective, patient-centered, timely, efficient, and equitable. Service-system outcomes are in turn necessary but not sufficient evidence of the ultimate client-level outcomes an initiative is meant to produce. A practice delivered with high fidelity, at real scale, sustained for years, can still fail to improve student well-being if the underlying practice was weak or mismatched to the population; the logic runs the other way too, a genuinely effective practice delivered with low fidelity in a handful of pilot classrooms will not show population-level effects no matter how good it is on paper. Neither tier substitutes for the other, and Proctor and colleagues were candid that the field's measurement toolkit for most of these eight constructs was still underdeveloped, existing instruments largely home-grown, and that measures built for efficacy trials would likely prove too cumbersome for real-world implementation studies, a gap that is part of why a platform-generated usage and fidelity signal, rather than a bespoke survey instrument, is a defensible design choice on its own terms.

That taxonomy gives BHPM a principled way to separate two kinds of indicator easy to conflate in a single reporting table. Fidelity, penetration, and adoption describe whether the system is being used as designed, by the people it was designed for, at meaningful scale, questions the platform can answer directly from its own usage records. Whether students are better off is a downstream, service-system and outcomes-level question the project's other components, the indicator set and the composite index, are built to address. Keeping the two tiers distinct, rather than collapsing "the platform was used" into "the platform worked," is a direct application of the Proctor framework, and it is the reasoning this section's evidence points to for treating process and outcome indicators as structurally separate families rather than variations on one scale.

Continuous Improvement Over Episodic Evaluation

A third strand concerns not just what gets measured but how measurement feeds back into practice while a program is still running. Deming's formulation of the Plan-Do-Study-Act cycle, developed for industrial quality management, offered a general method for testing a change on a small scale, studying what happened, and adjusting before committing to system-wide rollout **DEMING 1986**. A second contribution from that same body of work matters as much as PDSA itself: the distinction between special-cause and common-cause variation, whether a swing in a metric reflects something specific and addressable, or ordinary noise the underlying process always produces. Most fidelity- and process-monitoring plans lean on that distinction whether or not they name it, because treating every quarter-to-quarter wobble in a usage indicator as a signal worth acting on is how a monitoring system trains its own users to stop trusting it.

Bryk, Gomez, Grunow, and LeMahieu import that logic directly into K-12 improvement work, arguing that education reform tends to fail less

for lack of good ideas than for lack of disciplined methods to adapt those ideas to real, variable local conditions **BRYK ET AL. 2015**. Their book organizes the argument around six core principles: make the work problem-specific and user-centered rather than solving in the abstract; treat variation in performance across sites as the thing to understand, not an inconvenience to average away; see the whole system that produces current outcomes before changing a piece of it; embed measurement of outcomes and processes so the network can tell a genuine improvement from a coincidence; anchor practice change in disciplined PDSA inquiry rather than a single big-bang rollout; and accelerate improvement through networked communities, on the premise that a network learns faster collectively than any one site learns alone. Their proposed structure, the networked improvement community, organizes practitioners and researchers across sites around one specific, shared problem of practice with common measures, so that what one site learns through a PDSA cycle can inform the next site's cycle rather than staying local, a meaningfully different design than a generic professional learning community convened around a topic; the shared measures are what let the network compare notes and accumulate evidence across cycles rather than exchange anecdotes.

This is the direct literature basis for the Collaborative Learning Group structure specified under Activities 2.4 and 2.5 of the BHPM workplan. The quarterly learning-group cycle across the six participating LEAs reads naturally against the networked-improvement-community model rather than a status-meeting model: the composite index and indicator set are positioned to supply the common measures a network needs to compare notes across sites, and each quarter's summaries can function as the "study" phase closing the loop into the next quarter's "plan." The quarterly cycle earns that comparison only if it runs with the disciplined PDSA structure and shared measures Bryk et al. specify, not as a status-meeting series in new packaging; that is the standard the cycle should be checked against.

What The Evidence Says Works, Specifically, In School Mental-Health Implementation

A 2023 systematic review in Prevention Science narrows the general implementation-science literature to the specific setting BHPM operates in: universal, school-based mental-health prevention programming. Reviewing 21 studies published between January 2000 and October 2021, only five of them randomized controlled trials, the rest split across quasi-experimental, single-arm pre-post, cross-sectional, and qualitative or mixed-methods designs, the authors identified 22 distinct implementation strategies with some evidence of improving fidelity or adoption **BAFFSKY ET AL. 2023**. The strongest evidence concentrated in three places, and the review is specific about which outcome each one moved. Audit-and-feedback strategies, school or project staff observing program delivery and giving teachers direct, positive-reinforcement-oriented feedback, showed the most consistent effect on fidelity, positive results in four of five studies that tested it.

SECTION 05 · IMPLEMENTATION SCIENCE · CONTINUED

What The Evidence Says Works, Specifically, In School Mental-Health Implementation

Engaging school principals as visible program champions, local opinion leaders who promote the program and steer resources toward it, showed a consistent positive association with adoption specifically, in every qualitative study that examined it. Strategies building buy-in among front-line teachers touched both outcomes, fidelity in one study and adoption in two others.

The same review is candid about limitations worth naming rather than glossing over. Strategies aimed specifically at adoption, as distinct from fidelity, have been evaluated almost entirely through small qualitative or case-study designs, and the authors call explicitly for larger trials before treating those findings as settled. Fidelity, by contrast, was measured in sixteen of the twenty-one studies, adoption in only five, an imbalance any indicator set built from this review should carry forward honestly rather than implying adoption is as well-supported as fidelity is. The authors also flag that two-thirds of the reviewed studies had no explicit theoretical grounding for the strategy being tested, that most evidence comes from English-language, U.S.-based studies, and that staff turnover in schools introduces attrition problems that make longitudinal fidelity tracking harder to interpret over multi-year periods, exactly the kind of period a six-LEA initiative like BHPM runs across.

The practical implication for BHPM is narrower and more useful than a general endorsement of "monitoring." The strand of implementation strategy with the best evidence in this exact setting is continuous monitoring joined to feedback, not a periodic external audit, and that evidence is strongest specifically on fidelity rather than on adoption or the other six Proctor constructs. A system that surfaces usage and fidelity signals on a rolling basis, so a school or district can respond to its own implementation pattern between formal check-ins, is aligned with where this literature's evidence is strongest. A system that instead relies on an annual retrospective survey to reconstruct the prior year is aligned with a design this evidence base does not particularly support, and it compounds a second, well-documented

problem: self-report recall about program fidelity degrades over exactly the time horizon an annual instrument requires, the same demand-characteristic risk the Baffsky review flags in its own underlying studies.

How BHPM's Own Sequence Maps Onto These Frameworks

Layered on top of each other, these strands do not describe four separate literatures so much as one coherent pathway, and BHPM's own activity sequence follows it closely enough to state the mapping plainly. Deliverable 1's site readiness assessment corresponds to the exploration and installation stages in the Fixsen model, and, read through CFIR's inner-setting domain, to a check on organizational context rather than technical capacity alone, establishing before deployment whether a site has the preconditions the driver framework treats as necessary. Deliverable 2's configured deployment, the indicators, index, and scoring infrastructure this Brief documents, is the decision-support data system named directly as an organization driver. Deliverable 3's training and technical-assistance plan supplies the coaching component the Fixsen synthesis found training alone cannot substitute for. Deliverables 4 and 5's Collaborative Learning Group structure is the networked-improvement-community mechanism Bryk et al. specify, running on a PDSA cadence with the composite index as a candidate common measure. And underneath all four sits the platform's continuous analytics layer, which is what lets fidelity and adoption, in Proctor's terms, be observed as they happen rather than reconstructed once a year later, consistent with where Baffsky et al. locate the strongest evidence, specifically for fidelity, in this exact setting.

That correspondence is easier to see laid out stage by stage than carried in a single paragraph, so the diagram below puts the Fixsen model and BHPM's own sequence side by side.

Implementation sequence

The staged implementation sequence from readiness through learning groups. It is drawn as a diagram in the earlier working design.

[Open implementation sequence in the archive](#)

None of this substitutes implementation success for the separate, harder question of whether the initiative improves outcomes for students. Proctor's own three-tier logic is explicit that implementation outcomes are a necessary floor, not a proxy for the ceiling, and the Baffsky review's candor about its own evidence gaps, small samples, thin theoretical grounding, adoption limited to qualitative designs, is a reminder that even that floor is measured with

real uncertainty. What this literature does establish is that treating implementation as a measured, staged, continuously monitored process, rather than a launch event followed by an annual check-in, is itself an evidence-based design choice, distinct from any claim about student outcomes, and it is the choice BHPM's activity structure was built around.

SECTION 06 · COMPOSITE INDEX CONSTRUCTION

Composite Index Construction

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.1: how domain sub-indices roll up into a single composite score under the OECD/JRC composite-indicator construction framework: normalization, weighting, aggregation, banding and disclosure.

The Composite, As Activity 2.4 Fixes It



WHAT THE SECTION COVERS

- A Ten-Step Sequence, Not An Average
- Three Choices Carry The Most Consequence
- Two Further Commitments
- From Score To Color: Performance Banding

The full section, with every source beside the prose: bhp.mcare/deliverables/d2/index-construction

WHAT A SINGLE NUMBER TELLS A BOARD MEMBER

What does a single Behavioral Health Performance Index score actually tell a board member, and what does it hide? Picture two LEAs landing on the same composite number at the end of a reporting cycle. One runs high participation, strong process fidelity, and fast crisis response, but its outcome measures are barely moving. The other has thinner service utilization and a slower crisis-response track record, but real outcome gains wherever it is measuring at all. Averaged and compensated against each other inside a single index, those two different behavioral health systems can land on the same number, and a board reading only the top-line score has no way to tell them apart.

BOTTOM LINE · THE RECORD'S OWN SUMMARY

The BHPM is a constructed, published methodology, not a black-box score. Activity 2.4 fixes that construction as an actual composite rather than a framework left abstract: $BHPI = 0.40(\text{Voice}) + 0.20(\text{Behavior}) + 0.16(\text{Engagement}) + 0.12(\text{Academic}) + 0.12(\text{Service Receipt})$, reported inside Domain 4 (Outcomes & Impact) as that domain's flagship metric, not a fifth index competing with the Commission's four-domain taxonomy. Voice carries the heaviest weight because authentic student-voice evidence is the differentiating signal no other state instrument collects; Behavior is second because suspension, expulsion, and referral data is the highest-stakes signal already sitting in a district's SIS; Engagement (attendance) and Academic follow, with Academic weighted lightest by design since it is context for behavioral health rather than a driver of it; Service Receipt closes the loop and is optional and renormalizing. If a dimension is missing, the remaining weights scale up proportionally rather than penalizing a site for a data gap, the same 4-of-6 renormalization logic the student-level Wellness Index (.250–1.000, six BHPM-specific competencies: Self-Insight, Emotional Resilience, Relational Awareness, Conflict Resolution, Effective Help-Seeking, Reflective Growth) applies underneath BHPM's Voice dimension. Bands are Leading ($\geq .75$), Establishing ($\geq .60$), Developing ($\geq .45$), and Emerging ($< .45$). The Index Methodology page carries the full diagram, the omega commitment, and the small-cell suppression default. Every dashboard number carries the version marker of the methodology that produced it, and Activity 2.6's structured review with all six LEA teams settles the next one: the formula above is v0.9, under Commission and partner review.

SECTION 06 · COMPOSITE INDEX CONSTRUCTION · CONTINUED

Composite Index Construction

What does a single Behavioral Health Performance Index score actually tell a board member, and what does it hide? Picture two LEAs landing on the same composite number at the end of a reporting cycle. One runs high participation, strong process fidelity, and fast crisis response, but its outcome measures are barely moving. The other has thinner service utilization and a slower crisis-response track record, but real outcome gains wherever it is measuring at all. Averaged and compensated against each other inside a single index, those two different behavioral health systems can land on the same number, and a board reading only the top-line score has no way to tell them apart. That is not a flaw unique to behavioral health measurement; it is the problem composite-indicator methodology exists to solve, and why the Commission's four reporting domains (services provided, student demographics and characteristics, service utilization, and outcomes and impact) cannot simply be averaged into a single number without an explicit, published construction method behind the arithmetic.

A Ten-Step Sequence, Not An Average

Impacter Pathway's approach to the Behavioral Health Performance Index (BHPI) follows the ten-step composite-indicator framework set out by the OECD and the European Commission's Joint Research Center [OECD ET AL. 2008](#). The sequence begins well before any arithmetic runs: a theoretical framework tying each indicator to a construct the Commission actually asked for, deliberate variable selection rather than "whatever the platform already logs," and a documented, defensible treatment of missing data, so a partial-year deployment doesn't simply drop out of a domain sub-index and understate the sites furthest along the readiness curve. The framework also calls for multivariate checks (correlation structure and, where the indicator count supports it, a principal-components read) confirming indicators grouped into a domain actually move together; a domain that doesn't covary is measuring more than one construct under a single name. From there the sequence states a normalization method, a weighting method, and an aggregation method; requires sensitivity analysis on every prior choice; and closes with a presentation layer built for the audience that will actually read it: superintendents and county office leaders, not psychometricians. Each of the four Commission domains becomes a sub-index built through that sequence; the four sub-indices then aggregate into the BHPI itself. Nothing about the sequence is exotic. What is unusual, at least in K-12 behavioral health reporting, is publishing it in full rather than handing districts a single color and a shrug.

Three Choices Carry The Most Consequence

Three of those ten steps carry the most consequence and the most disagreement in the literature, which is why the full Index Methodology page treats each at length rather than summarizing it away here. Normalization has no neutral option. Z-scores center every indicator on the sample's own mean and spread, min-max

rescaling stretches each indicator across a fixed range anchored to the observed minimum and maximum, and distance-to-target methods measure a site's distance from a policy-set benchmark instead of how its peers happened to perform that cycle, three different implicit questions about what "average" means, and with a program built on six LEAs and eighteen schools, peer-relative normalization is also the most exposed to noise from a small comparison set. The choice changes which schools look strong before a single weight is assigned.

Weighting is the step most exposed to hidden values, and the framework catalogs several families rather than endorsing one. Equal weighting looks neutral but quietly asserts every indicator in a domain matters the same amount, itself a judgment, just an unexamined one. Data-driven weighting through principal components assigns weight by how much variance an indicator shares with its domain, surfacing structure equal weighting would flatten, but it can swing sharply with a handful of added or dropped sites in a program this size. Budget-allocation or expert-elicitation weighting (the family associated with structured stakeholder input, and the natural fit for the crosswalk's Activity 2.6 LEA-team consultations) asks reviewers to directly allocate a fixed number of importance points across indicators, making the value judgment visible and arguable instead of laundering it through statistics. Visible and arguable is preferable even when less tidy: a weighting scheme nobody can see is one nobody can contest.

This is where the sequence stops being a formality and starts deciding things on the Commission's behalf. Aggregation decides whether a strength in one domain can offset a weakness in another. Linear, compensatory aggregation, a straight weighted average, lets exactly the scenario that opened this section happen: strong process fidelity and fast crisis response mathematically canceling out flat outcomes, or the reverse, with the resulting number giving no indication which trade occurred. Geometric or otherwise non-compensatory aggregation requires every domain to clear its own floor before it can contribute fully to the composite, real stakes for a behavioral health index specifically, where a district should not be able to average away a crisis-response gap with strong attendance numbers, and a non-compensatory rule is the mechanism that prevents it. The Index Methodology page works through both families with the sensitivity analysis [OECD ET AL. 2008](#) requires (Monte Carlo-style resampling across plausible combinations of normalization, weighting, and aggregation choices, testing whether a site's ranking holds or swings depending on which defensible specification was picked) and lays out BHPI's actual, Commission-reviewed construction under Activity 2.4: a linear weighted sum of five components (Voice 0.40, Behavior 0.20, Engagement 0.16, Academic 0.12, Service Receipt 0.12), reported as Domain 4's flagship outcome metric rather than as a hypothetical.

Two Further Commitments

Two further commitments round out the construction and are non-negotiable regardless of which weighting or aggregation choice the Activity 2.6 structured review lands on. First, sub-index and composite reliability will be reported using coefficient omega rather than Cronbach's alpha.

Two Further Commitments

Alpha assumes tau-equivalence, that every indicator feeding a scale carries the same relationship to the underlying construct, and each BHPI domain's indicator set violates that assumption by design, mixing process counts, participation rates, and outcome measures with different units and different loadings on whatever the domain is measuring. Omega is derived instead from a factor-analytic model that lets each indicator's contribution to reliability track its actual loading rather than assuming a uniform one, and it stays defensible under exactly the heterogeneous-indicator condition where alpha is known to mislead (MCDONALD 1999); both a domain sub-index and the composite it feeds can be estimated this way, not only the top level.

Second, the public-facing dashboard applies small-cell suppression by default, following the disclosure-avoidance practice federal longitudinal-data guidance recommends for any subgroup report where re-identification risk rises as group size falls (SEASTROM 2010). That guidance distinguishes primary suppression, masking any cell below the minimum group size on its own, from complementary suppression, which masks additional cells in the same table that would otherwise let a reader back-calculate the suppressed value by subtraction from a published total. It documents states setting that threshold anywhere from five to thirty students, ten the most common, and recommends one stated rule rather than deciding case by case; the default here is ten, set with SCCOE data governance. That risk isn't abstract in an eighteen-school program: the smallest cells usually sit at the intersection of a small school and a small

demographic subgroup, exactly what complementary suppression exists to catch.

From Score To Color: Performance Banding

For performance banding, translating a continuous index score into something a board can act on, the Index Methodology page proposes adapting the in-state precedent every California LEA already knows: the School Dashboard's five-by-five status-and-change grid (CALIFORNIA DEPARTMENT OF EDUCATION. 2025). Status reflects a percentile-ranked performance level in the current year; Change reflects year-over-year movement; crossing the two five-level scales produces twenty-five cells mapped onto a five-color scale. That design already generalizes across dissimilar indicator types on the Dashboard (chronic absenteeism, suspension rate, graduation rate, and academic indicators all run through the same logic), evidence the grid was not built for one narrow use case, and its technical guide is archived as a new, separately versioned document every year rather than silently revised in place, a discipline the BHPI's own v0.9 version marker is modeled on. Reusing the grid means districts read the BHPI the way they already read academic and climate indicators, rather than learning a second accountability grammar from scratch, and it lets the two hypothetical districts from the opening actually separate: one lands in a cell showing high status but flat change, the other lower status but strong positive movement, giving a board two distinguishable pictures instead of one blurred average.

SOURCES CITED IN THIS SECTION · 4	
California Department of Education, 2025 The five-by-five grid crosses a percentile-based Status level against a year-over-year Change level, producing 25 color-coded cells across five performance colors (Blue, Green, Yellow, Orange, Red).	McDonald, 1999 Derives coefficient omega from a factor-analytic measurement model rather than a tau-equivalent one.
OECD et al., 2008 Establishes a ten-step sequence for constructing a defensible composite indicator, treating each step (normalization method, weighting method, aggregation rule) as a substantive methodological choice rather than a mechanical one.	Seastrom, 2010 Most states set a minimum reporting-group size between 5 and 30 students, with 10 the most common (modal) threshold, specifically to prevent indirect re-identification of individual students.

SECTION 07 · RESEARCH LINEAGE

The Impacter Pathway Research Lineage

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.1: the peer-reviewed and technical lineage behind the measurement and scoring approach the program brings to the cohort.

The Lineage, Traced End To End

1969

Bivariate Granger Causality

Granger · Econometrica

Testable, operational definitions of causality and feedback between two time series, decomposing their cross-spectrum into causal-lag and causal-strength measures.

2007

Lasso-Granger: Multivariate Graphical Granger

Arnold, Liu and Abe · ACM SIGKDD

Extends the bivariate framework to many variables at once, using Lasso-based variable selection to recover sparse causal graphs. First author Dr. Andrew Arnold is now Impacter Pathway's Senior Scientific Advisor.

2026

Impacter Model Suite v6.0

Production scoring pipeline · Model Card

DeBERTa-v3-base (184M parameters) with a CORAL ordinal-regression head, architected under the same causal-inference discipline. Average QWK 0.918, 77.3% exact match, 98.0% adjacent agreement across eight competencies.

Year 2

Stanford Medicine Upstream-Lever And Dosage Analysis

Joint with Stanford Medicine, once the indicator set stabilizes

The same multivariate causal-inference toolkit applied to ask whether an upstream lever, a training rollout, a screening-cadence change, a service-dosage shift, plausibly caused a downstream change in a student outcome trajectory.

WHAT THE SECTION COVERS

- The Causal-Inference Lineage Behind The Scoring Pipeline
- Continuous Calibration In Production
- The Construct-Coding Evidence For A Deterministic Architecture
- The Lineage, Traced End To End
- Reflection As Mechanism, And One Deliberate Boundary

The full section, with every source beside the prose: bhpm.care/deliverables/d2/research-lineage

BOTTOM LINE · THE RECORD'S OWN SUMMARY

Impacter Pathway's scoring architecture is a lineage, not a black box. A named causal-inference researcher, credentialed by a specific and checkable methodological line rather than a generic "machine learning" bio, leads it as Senior Scientific Advisor and pipeline architect. A documented weekly human-in-the-loop calibration process stratifies review across construct, score band, prompt, and item type. And the scoring design itself is discriminative, deterministic, and non-generative, a choice the current construct-coding literature (Ober et al., 2026) directly supports over generative alternatives. That lineage, not the model card's accuracy figures on their own, is what makes the pipeline trustworthy. One boundary holds regardless of how this framing gets reused, edited, or reorganized elsewhere: the Saab et al. (2009) citation is founding neurobiological inspiration only, never screener validity evidence, a distinction load-bearing for how a clinical or research reviewer reads the whole citation set.

SECTION 07 · RESEARCH LINEAGE · CONTINUED

The Impacter Pathway Research Lineage

Every algorithmic system carries an intellectual genealogy, whether a partner ever asks about it. For a scoring pipeline sitting inside a public-sector behavioral health contract, that genealogy is not incidental biography. It is evidence. Reviewers assessing a machine learning system for defensibility want to know two things that get conflated in casual descriptions of "the technology": who built the statistical judgment embedded in the model, and where the measurement philosophy behind the model's design choices actually comes from. This section answers both, tracing Impacter Pathway's scoring architecture back to a specific, checkable research lineage rather than treating the pipeline as a black box that arrived fully formed. That lineage runs in five distinguishable threads, and this section keeps them separate on purpose rather than blending them into one undifferentiated "the research supports us" claim: who architected the scoring pipeline and what causal-inference toolkit they bring to it; how that pipeline is kept honest in production between formal reviews; what the current construct-coding literature says about the specific architectural choice Impacter Pathway made; why the prompts ask students to reflect rather than rate themselves; and, stated as its own explicit boundary rather than folded quietly into the others, what a piece of mouse neuroscience is and is not doing in this citation set.

The Causal-Inference Lineage Behind The Scoring Pipeline

Impacter Pathway's Senior Scientific Advisor, Dr. Andrew Arnold, is the architect of the platform's scoring pipeline, per the organization's published Model Card. His public professional record runs through machine learning research and applied engineering at a scale relevant to a system like BHPM's: Principal Machine Learning Engineer at Shopify, former Chief Scientist for AWS CodeWhisperer, and a PhD in Machine Learning from Carnegie Mellon University. Before those industry roles, his academic work concentrated on a narrower and more technical problem than "machine learning" as a category suggests: recovering causal structure from time-series data. His 2007 paper with Yan Liu and Naoki Abe, presented at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, introduced graphical Granger-causality methods for multivariate time series (ARNOLD ET AL. 2007), formal techniques for asking, given several variables observed over time, which ones plausibly drive which. The paper's central contribution, the Lasso-Granger algorithm, uses Lasso-based variable selection to recover sparse causal graphs among many time-varying variables simultaneously, and the authors characterize the relative performance of a family of graphical Granger approaches from multiple angles, showing that the Lasso-based method gives consistent gains over the standard pairwise graphical Granger approach for that recovery task.

That paper does not stand on its own; it extends a specific, older methodological lineage rather than inventing the underlying idea of

causal inference from scratch, and naming that lineage is part of what makes the credential checkable rather than decorative. The bivariate notion of Granger causality traces to Clive Granger's 1969 paper in *Econometrica*, which proposed testable, operational definitions of causality and feedback between two time series, showing that the cross-spectrum between them decomposes into components attributable to distinct causal arms, from which a measure of causal lag and causal strength could be constructed (GRANGER 1969). Granger's original framework handled one pair of variables at a time. What Arnold, Liu, and Abe's 2007 paper adds is the move from that bivariate setting to a multivariate, graph-structured one: instead of asking whether variable A Granger-causes variable B in isolation, Lasso-Granger asks which sparse subset of many candidate variables plausibly drives each other variable, using variable-selection machinery to keep the resulting causal graph interpretable rather than a dense, uninformative tangle of pairwise arrows. This is a specific, load-bearing kind of expertise for a program like BHPM, and the fit is not decorative. Activities under this contract will eventually need to ask a harder question than whether an indicator moved. They will need to ask whether a given upstream lever (a training rollout, a change in screening cadence, a shift in service dosage) plausibly caused a downstream change in a student outcome trajectory, as opposed to merely preceding it on a calendar, across a system with many moving indicators rather than one clean pair of variables. That is precisely the class of problem graphical Granger methods were built to address, and it is the planned analytic foundation for the Year 2 upstream-lever and dosage analysis Impacter Pathway intends to run jointly with Stanford Medicine once the indicator set stabilizes. The KDD 2007 paper and the 1969 paper behind it are cited here for that reason: not as evidence about the behavioral health screener itself, but as the methodological throughline connecting the pipeline's architect, through a named and dated origin point in the causal-inference literature, to the specific statistical toolkit this contract will eventually need for causal, not merely correlational, questions.

Continuous Calibration In Production

The production expression of that same discipline shows up closer to the ground, in how Impacter Pathway keeps its scoring models honest between formal review cycles, and the mechanics of that process are more specific than a one-line description of "human review" suggests. The organization's internal technical note on the scoring pipeline's calibration layer describes a weekly, stratified, human-in-the-loop sampling process (IMPACTER EDUCATION CORP. 2026) built around four independent stratification factors rather than a single random draw. Competency is sampled on its own axis so that no rubric-aligned construct goes more than one cycle without review, because agreement statistics behave differently across constructs of varying operational clarity and aggregating across them can mask construct-specific drift, the same lesson the construct-coding literature discussed below documents from a different angle.

SECTION 07 · RESEARCH LINEAGE · CONTINUED

Continuous Calibration In Production

Predicted score band is a second axis, and it is deliberately not sampled uniformly: the clearest low- and high-evidence responses tend to be the most stable part of any rubric, while the middle bands are where neighboring levels are hardest to distinguish and where calibration errors are most likely to hide, so the sampling layer intentionally over-represents those middle bands rather than letting easy extreme cases dominate the review set and inflate agreement statistics. Course or prompt family is a third axis, protecting against prompt-level overfitting and ensuring newly authored content gets reviewed within one sampling cycle of release rather than waiting for a future audit to notice it. Item and response type (voice-elicited versus text-elicited, short-form versus long-form) is the fourth, because automatic speech recognition introduces a known source of input-level noise the technical note treats as something to represent explicitly in the validation distribution rather than average away. Responses selected this way surface inside a reviewer-facing interface where an expert either confirms or overrides the model's predicted score against rubric anchor language, and every re-scored item is written back with full provenance (the original prompt, the model version that produced the score, the reviewer's rating, and a timestamp) so that a later disagreement can be traced to a model-version change, a human re-rating, or genuine rubric ambiguity rather than treated as unexplained noise. Re-scored items then flow automatically into two places: a rolling recomputation of agreement statistics against the most recent human ratings, with movement beyond a pre-registered degradation threshold flagging a competency for targeted retraining or rubric re-examination, and the training partition for the next scheduled fine-tune. Model promotion itself stays gated the same way regardless of how the new training data was sourced: a candidate model is only promoted to production if its held-out agreement with human raters improves and clears the organization's standing thresholds, and once promoted it is frozen for deterministic inference (IMPACTER EDUCATION CORP. 2026). This is what human-in-the-loop review means in practice for this system: not a one-time human check performed before launch, but a standing sampling design that keeps expert raters continuously embedded in what the model is permitted to learn next, with the sampling logic itself doing real methodological work rather than functioning as a compliance gesture.

The Construct-Coding Evidence For A Deterministic Architecture

That same commitment to construct clarity over raw model scale connects Impacter Pathway's architecture to a second, independent research thread with its own institutional pedigree, and this is the thread where the design choice and the evidence for it line up most directly. Ober, Zhang, Zapata-Rivera, Schroeder, and Botelho's 2026 study in the Journal of Educational Data Mining compared multiple ChatGPT-4o configurations, varying both temperature and model size, against expert human coders on more than 10,000 conversational turns from 107 English learners (OBER ET AL. 2026). The coding task spanned several constructs drawn from a conversation-

based math assessment with a pedagogical agent (persistence, self-efficacy, interest, perceived challenge, and prior knowledge), and the study's central finding was not that a better or larger model solves construct coding. It was that how precisely a construct is operationally defined governs coding reliability far more than which model, or what model size, is doing the coding. Self-efficacy, the most operationally well-specified construct in the study, showed the strongest alignment between human and machine coders; more theoretically diffuse constructs like persistence did not, and reliability for those diffuse constructs varied unpredictably depending on the temperature setting used, in a pattern the authors could not fully explain by construct alone. The paper documents genuine non-determinism and reproducibility problems when generative language models are asked to code the same psychological construct twice under different sampling conditions, the same instability, arrived at from a different research tradition, that Turpin, Michael, Perez, and Bowman documented in generative chain-of-thought reasoning more broadly, where models can produce fluent explanations that systematically misrepresent the actual basis for their own output (TURPIN ET AL. 2023). That finding sits directly behind an architectural decision Impacter Pathway made and has not revisited: its production scoring models are discriminative and ordinal, not generative. A fine-tuned, 184-million-parameter DeBERTa-v3-base encoder paired with a CORAL ordinal-regression head (CAO ET AL. 2020) produces one deterministic score per response under a frozen, versioned model, evaluated across the eight competencies at an average quadratic-weighted kappa of 0.918, 77.3% exact-match agreement with human raters, and 98.0% adjacent agreement (IMPACTER EDUCATION CORP. 2026). It does not sample text on inference, and it does not produce a different judgment on a second pass over the same response. The instability Ober and colleagues document, and that Turpin and colleagues document in a related setting, is a property of asking a generative model to code a construct by sampling; it is not a property Impacter Pathway's scoring architecture inherits, because the architecture was built specifically to avoid that failure mode by never sampling in the first place. The organization's partner-facing guide to the screening approach describes the same design choice in plainer terms, and frames it against a broader pattern in validated assessment rather than as a bespoke Impacter Pathway invention: students respond to open-ended prompts in authentic voice, and a rubric-aligned scoring engine trained against expert human raters produces the developmental score that routes a student within a multi-tiered system (IMPACTER EDUCATION CORP. 2026). The guide's own framing is worth carrying forward here because it locates Impacter Pathway inside a familiar category rather than outside it: assessments that standardize the rubric rather than the prompt (AP exams, NAEP writing, ETS performance items, Smarter Balanced constructed-response components) accept that different respondents see different stimuli and place reliability in the scoring system instead, exactly the pattern Impacter Pathway's architecture instantiates for behavioral health screening.

SECTION 07 · RESEARCH LINEAGE

The Lineage, Traced End To End

Traced end to end rather than argued paragraph by paragraph, that causal-inference lineage forms a single visual line: a bivariate econometric framework in 1969, its multivariate extension at KDD in

2007, the production scoring pipeline that same discipline now leads, and the Year 2 causal analysis it is built to eventually support.

Research lineage timeline

The dated sequence of published work behind the scoring approach. It is drawn as a diagram in the earlier working design.

[Open research lineage timeline in the archive](#)

Reflection As Mechanism, And One Deliberate Boundary

Two further citations round out this lineage, and the framing on both is deliberately narrow, by design. Hirshberg, Colaianne, and colleagues' multi-university study of a for-credit undergraduate flourishing course found that students enrolled in a course pairing academic content with structured experiential and reflective practice showed improved flourishing and mental-health outcomes relative to comparison students who were not enrolled **HIRSHBERG ET AL. 2022**. That study did not originate the idea that structured reflection does real psychological work rather than merely delivering information; it extends a research line that runs back at least to Pennebaker and Beall's 1986 study establishing what later became known as the expressive-writing paradigm, in which students who wrote about a personally traumatic event on four consecutive days, rather than about a trivial topic, showed a short-term rise in physiological arousal and negative mood followed by measurably fewer health-center visits over the following six months **PENNEBAKER ET AL. 1986**. The specific finding in each of these two papers is different (one is about writing and subsequent physical-health utilization in a general undergraduate population, the other about a structured course and flourishing outcomes), but the throughline connecting them is the same premise Hirshberg's study tests directly and Impacter Pathway's prompt design assumes: that reflecting on one's own experience in one's own words is itself a mechanism, not simply a format for delivering a score to a waiting rubric. That premise belongs in this section as design rationale, not as validity evidence for a screener: it is part of why Impacter Pathway's prompts ask students to reflect in their own words rather than select from a fixed checklist, on the premise that structured reflection is itself a meaningful unit of measurement.

The second citation requires an explicit boundary, stated plainly rather than implied. Saab and colleagues' 2009 study in Neuron found that inducible overexpression of a calcium-sensor protein, NCS-1, in the dentate gyrus of adult mice selectively increased exploratory behavior alongside facilitated synaptic plasticity, including facilitated long-term potentiation, and faster spatial-memory acquisition, an effect reversible with a peptide that blocked

the protein's binding to the dopamine D2 receptor **SAAB ET AL. 2009**. The manipulation was inducible and its effect was pharmacologically reversible, which is what let the authors attribute the behavioral change specifically to that protein interaction rather than to some earlier developmental artifact of the mouse's genetics, a methodologically careful study, on its own terms. But its own terms are mouse neurobiology. It appears in Impacter Pathway's research lineage as founding inspiration for taking seriously the biological basis of exploratory behavior as a construct worth measuring in young people at all, nothing more. It is not evidence that the screener is valid, it is not evidence about human behavioral health outcomes, and it should not be read as either. Mouse neuroscience does not validate a human skills measurement model, and this section states that distinction directly so that no reader has to infer it.

Set side by side, what these seven citations describe is an organization whose measurement choices trace back to identifiable, checkable sources rather than internal intuition alone: a causal-inference researcher, credentialed by a named and dated methodological lineage running from Granger's 1969 bivariate framework through his own 2007 multivariate extension, leading the scoring architecture; a documented weekly human-calibration process, stratified across four independent factors rather than a one-time validation event; a published construct-coding literature that independently supports the deterministic-over-generative design choice Impacter Pathway already made, and that Impacter Pathway's own architecture happens to sidestep by construction rather than by claim; a reflective-practice research base running from a 1986 expressive-writing study through a 2022 multi-university flourishing course behind the prompt design itself; and one explicitly bounded piece of mouse neuroscience that this section refuses to let drift into a claim it was never built to support. The organization's own plain-language guide to its machine learning model states the operating principle underneath all of it directly: human judgment, expressed through rubrics and expert raters, leads the system, and machine learning scales that judgment rather than replacing it **IMPACTER EDUCATION CORP. 2026**. That is the claim this section exists to support with citations rather than assertion.

SECTION 08 · INDEX METHODOLOGY

The Behavioral Health Performance Index · Construction Methodology

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.4: define the Behavioral Health Performance Index: the weighted formula, the missing-data renormalization rule, the four performance bands, and the per-student rubric beneath the Voice dimension.

The Working Formula, Version 0.9

Voice 40%	Behavior 20%	Engagement 16%	Academic 12%	Service Receipt 12%
DIMENSION	WEIGHT	WHY IT CARRIES THAT WEIGHT		
Voice	0.40	Weighted heaviest because authentic student-voice evidence, what a student actually says in their own words, is the differentiating signal no other state behavioral-health instrument collects.		
Behavior	0.20	Second-heaviest because suspension, expulsion, and referral data is the highest-stakes signal already sitting in a district's own student information system.		
Engagement	0.16	Attendance is a strong, well-established wellbeing signal: present, but weighted below Voice and Behavior rather than treated as equally decisive.		
Academic	0.12	Lightest by design: academic performance is context for behavioral health, not a driver of it, so it contributes without dominating the composite.		
Service Receipt	0.12	Closes the loop between what was delivered and what it produced. Optional and renormalizing, so a site without service-receipt data yet is not penalized for it.		

FROM SCORE TO COLOR · FOUR BANDS ON THE 0 TO 1 AXIS

Emerging < .45	Developing ≥ .45	Establishing ≥ .60	Leading ≥ .75
-------------------	---------------------	-----------------------	------------------

If a site is missing data for one of the five dimensions, that dimension is excluded and the remaining weights scale up proportionally so they still sum to 1.0. A site is never penalized for a data gap, it is scored on what it has actually reported.

WHAT THE SECTION COVERS

- From Four Domains To One Index
- Building The Sub-Indices: Selection, Imputation, Coherence
- Normalization: Choosing What "Comparable" Means
- Weighting: Making The Value Judgments Visible
- Aggregation And The Compensability Question
- From Score To Color: Performance Banding
- Composite Reliability: Omega, Not Alpha
- Uncertainty Quantification And The Mandatory Sensitivity Check
- Protecting Students: Suppression And The No-Student-Level-Data Rule
- Keeping The Method Honest: Versioning And Public Documentation

The full section, with every source beside the prose: bhpm.care/deliverables/d2/index-methodology

BOTTOM LINE · THE RECORD'S OWN SUMMARY

BHPI never stands alone as a single fixed number. The version marker, not the score by itself, is the citable unit: any reported figure carries its version (v0.9), the real formula and per-dimension rationale documented above, the missing-data renormalization rule, and the four performance bands together, so a reader can see exactly what produced a district's score, and why Voice carries the heaviest weight in that sum.

Performance bands read in a plain-language, strengths-based grammar: either BHPI's own four-band scale (Leading, Establishing, Developing, Emerging) or the California Dashboard's status-and-change grid described above, so LEA teams read behavioral health performance in a register they already recognize. Which of the two the production dashboard leads with, or whether the two get reconciled into one, is Activity 2.6's call. Any breakdown by demographic or small subgroup pairs with the minimum-cell-size suppression rule described above, once that rule has SCCOE data-governance sign-off.

Composite reliability (omega) and scoring-model agreement (QWK) stay visibly separate figures. They answer different questions, and conflating them overstates what either one actually certifies.

The real formula, the bands, and the renormalization rule above are settled at v0.9. Activity 2.6's structured review with all six LEA teams sets the sensitivity analysis and the final banding.

SECTION 08 · INDEX METHODOLOGY · CONTINUED

The Per-Student Rubric Beneath Voice: Six Competencies, Four Levels

These six competencies were developed specifically for BHPM, co-developed with the San Diego County Office of Education, a distinct, BHPM-specific set, not the general IMPACTER 8 competency model used elsewhere on the platform. Every student needs at least 4 of the 6 competencies reported to receive a Wellness Index score. If one or two are unreported, the index renormalizes across the remaining reported competencies, the same proportional-scaling logic BHPM itself uses at the site level.

COMPETENCY	1 · FOUNDATIONAL	2 · DEVELOPING	3 · COMPETENT	4 · ADVANCED
Self-Insight	Struggles to notice or describe their actions, motives, or values; avoids self-reflection or misreads how actions connect to personal identity; little sense of learning from experience.	Sometimes recognizes the gap between actions and values but is vague or needs prompting; limited detail about what was realized or how self-insight could lead to growth.	Clearly describes a moment when actions didn't fit their values; explains what was realized, how self-understanding changed, and gives examples of learning from experience.	Offers deep, thoughtful insight into motives, identity, and growth; connects actions, values, and learning; reflects on how increased self-insight guides future choices and helps others.
Emotional Resilience	Has difficulty describing how they manage or recover from difficult emotions or stressful situations; may focus on the problem without naming feelings or using any coping strategies.	Sometimes describes handling tough moments or feelings but gives only general or inconsistent details; may mention strategies or supports but lacks depth or clarity.	Clearly explains how they responded to and managed a challenging emotion or situation; names specific coping strategies, supports, or resources used to recover or move forward.	Shares rich, specific examples of managing adversity and emotions; explains how resilience developed over time; reflects on using and adapting strategies to recover, grow, and help others facing challenges.
Relational Awareness	Rarely notices or describes changes in relationships or their own role; may focus on disconnection, conflict, or confusion; shows little empathy or reflection on impact or learning.	Sometimes recognizes changes or shifts in relationships and can mention feelings or outcomes, but responses lack detail or personal insight and often require prompting.	Clearly describes a relationship change, their role in it, and how it felt; explains what was learned about themselves or others, and how they adapted or grew from the experience.	Offers nuanced, thoughtful insight into relationship dynamics and shifts; reflects on personal and others' perspectives; discusses how relational awareness shapes trust, belonging, and future relationships.
Conflict Resolution	Avoids or minimizes discussion of conflict or making amends; focuses only on the problem or harm; does not mention repair or solutions; little reflection on role or outcomes.	Sometimes discusses trying to resolve conflict or repair relationships, but with limited details; needs prompting to explain own actions, impact, or learning from the experience.	Describes actions taken to resolve conflict, repair harm, or make things right; explains steps clearly, what was learned, and reflects on personal growth in conflict situations.	Explains conflict resolution deeply; describes leading or modeling repair for others, using restorative practices, and promoting positive change in relationships and the community.
Effective Help-Seeking	Reluctant or unable to talk about asking for help; may express distrust, shame, or discomfort using supports; rarely explains benefit or value of seeking help.	Sometimes describes seeking help, but inconsistently or with uncertainty; benefits of support may be unclear or responses lack detail; may need prompting to discuss using resources.	Clearly explains when and how they sought help or support; describes how seeking help made a positive difference for themselves or others; reflects on value of support systems.	Regularly seeks help confidently; explains how support benefits self and community; encourages others to seek help and describes effective strategies for using resources in spoken reflection.
Reflective Growth	Struggles to notice or describe personal change; may feel stuck or unable to reflect on growth, learning, or how experiences have shaped their identity or choices.	Sometimes mentions changes in themselves or growth, but descriptions are vague, general, or require prompting to connect change to specific actions or learning.	Clearly explains a meaningful way they have changed; describes how growth happened, what contributed to it, and what was learned or how perspective shifted through reflection.	Gives rich, detailed reflection on their growth over time; connects changes to actions, learning, and experience; explains how reflective growth informs ongoing improvement and helps others believe in their own growth.

SECTION 08 · INDEX METHODOLOGY · CONTINUED

The Behavioral Health Performance Index · Construction Methodology

Version

Construction methodology v0.9, for Commission and partner review · weighting and banding finalize through the Activity 2.6 structured review with all six LEA teams.

A single number invites a single question: how was it built? A second question follows close behind: could someone who was not in the room reconstruct it from what's published, line by line, without having to take anyone's word for it? The Behavioral Health Performance Index (BHPI) exists to give the Commission, county office leadership, and the six participating LEAs one defensible read on behavioral health performance across a school year, and answering both questions in the open is the actual work of this page. Not a summary for a board slide, but the working method an outside reviewer, an auditor, or a skeptical superintendent could rebuild from scratch, decision by decision, including every judgment call inside them. It applies the ten-step construction framework set out jointly by the OECD and the European Commission's Joint Research Center (OECD ET AL. 2008) to the four reporting domains the Commission specified for BHPM: services provided, student demographics and characteristics, service utilization, and outcomes and impact. Every section below names which of those ten steps it is answering, states the choice BHPI v0.9 makes, and names where the Activity 2.6 review settles it.

From Four Domains To One Index

Step one of the OECD/JRC framework is a theoretical framework: naming, before any data collection, what each indicator is supposed to represent and how it ties back to a construct someone actually asked for (OECD ET AL. 2008). For BHPI, that construct is set by the Commission's four domains rather than by whatever the platform happens to already log. Each domain becomes its own sub-index (a Services Sub-Index, a Demographics/Characteristics Sub-Index, a Utilization Sub-Index, and an Outcomes/Impact Sub-Index) built independently through the full ten-step sequence before the four sub-indices are combined into the single BHPI score a district sees on the public dashboard. Building sub-indices first and combining second, rather than blending every indicator into one flat weighted average, keeps the domain structure legible: a reader can see which of the four Commission domains is driving a district's overall standing, not just the standing itself. It also means a data quality problem or a contested weighting choice in one domain stays contained to that domain's sub-index instead of silently diffusing across the whole score in a way no one downstream could trace back to its source.

Within each domain, the candidate indicator set comes from the crosswalk developed under Activity 2.3, the same indicator families documented on the S4 page (process, outcome, and contextual indicators), rather than from a separate index-specific indicator hunt. That continuity matters for defensibility: a reviewer checking the index methodology should find the same indicators, defined the

same way, as a reviewer checking the Metrics Brief itself. It also means the two documents cannot drift apart over time without the drift being visible; if a future revision adds an indicator to the crosswalk without a matching update here, or vice versa, that mismatch is itself a signal that one of the two documents needs to catch up to the other, not a permissible state to leave standing.

Building The Sub-Indices: Selection, Imputation, Coherence

Steps two through four of the OECD/JRC sequence are the unglamorous groundwork that most published composites skip past, and BHPI does not get to skip them because the Commission's audience will ask (OECD ET AL. 2008). Variable selection has to be documented well enough that a reader can see why an indicator is in the index and not simply available in the platform; an indicator earns its place in a sub-index because it maps to a construct the Commission's domain language actually names, not because a report happened to already compute it. Missing-data treatment has to be stated in advance: which imputation rule applies when a school has not yet reported a given quarter's data, and when a cell is instead suppressed rather than imputed (the suppression rule is covered separately below, since it is a privacy control as much as a statistical one). A quarter with no data and a quarter with data below the suppression threshold are two different situations that call for two different labels on the dashboard, and conflating them, treating an unreported quarter as if it were a reported-but-suppressed one, would misstate what the platform actually knows about a district in that period.

Before any indicator is normalized or weighted, the indicators inside a domain need a coherence check: do they actually move together, in the way that justifies treating them as facets of one underlying sub-index, or is a domain quietly bundling two unrelated constructs under one label? The OECD/JRC handbook recommends multivariate methods (correlation structure and, where the indicator count supports it, factor analysis) as that check, and flags a domain that fails it as a candidate for splitting rather than forcing it into a single sub-index (OECD ET AL. 2008). This is not a formality that gets run once and filed away. A domain whose indicators pass the coherence check in year one of BHPM can fail it in year two, once six LEAs of different size, staffing model, and program maturity start producing genuinely different indicator profiles from one another; a services-domain sub-index that looked like one coherent construct when every site was still ramping up implementation could split into a training-completion facet and a service-delivery facet once sites diverge in how far along they are. BHPM applies that check per domain before finalizing sub-index composition, and will report the result of that check, and re-run it at each subsequent Activity 2.6-style review, alongside the sub-index definitions once the Activity 2.6 review closes.

SECTION 08 · INDEX METHODOLOGY

Normalization: Choosing What "Comparable" Means

Indicators arrive in incompatible units (a completion rate, a days-to-service count, a tier-movement percentage, a per-pupil dollar figure), and none of them can be added together until they are put on a common scale. The OECD/JRC handbook treats normalization as a substantive methodological choice rather than a mechanical step, because each common option answers a different implicit question (OECD ET AL. 2008). Z-scoring asks "how many standard deviations from the six-LEA mean," which is comparable and interpretable but sensitive to outliers and to the specific comparison group in view; with only six LEAs contributing to that mean, one site's unusual quarter has more leverage over everyone else's z-score than it would in a larger reference population, a distortion worth naming plainly rather than treating as a footnote. Min-max rescaling asks "where does this fall between the observed floor and ceiling," which is bounded and intuitive but means a single unusually strong or weak district shifts everyone else's scale, and with six LEAs that ceiling or floor can be set by a single site's single quarter. Distance-to-target normalization asks a different question entirely: "how far from a stated goal," which anchors the index to policy targets rather than to relative performance but requires the Commission to set defensible targets for every indicator before the index can run, and a target set carelessly, too easy or too aspirational relative to what the six sites can realistically reach in the contract period, would distort the resulting score just as badly as an unstable comparison group would.

At v0.9, the sub-indices use distance-to-target normalization where the Commission's crosswalk already implies a target (for example, indicators tied to a completion or fidelity threshold) and z-scoring elsewhere, pending the full indicator-by-indicator normalization table the Activity 2.6 review will finalize. The rationale is that a behavioral health performance measure serving six LEAs of different size and baseline capacity should not let a single high- or low-performing site distort everyone else's normalized scale the way min-max rescaling would, and should use fixed policy targets wherever the underlying program (a fidelity checklist, a service-completion standard) already defines one. That rationale is a working position, not a settled one: the specific list of which indicators get a distance-to-target treatment versus a z-score treatment is exactly the kind of indicator-by-indicator call the Activity 2.6 review exists to confirm or revise, and this page will publish the finalized table, not just the general rule, once that review closes.

Weighting: Making The Value Judgments Visible

Every weighting scheme embeds a value judgment about which indicators matter more; the only real choice is whether that judgment stays implicit or gets stated and defended (OECD ET AL. 2008). Equal weighting is the most common default in published composites and the easiest to explain, but it quietly asserts that every indicator in a domain deserves identical influence, an assertion nobody actually made on purpose; it is a default chosen by not choosing. Statistical weighting via principal components analysis lets the data itself set weights based on shared variance, which removes the appearance of an arbitrary human choice but replaces it with an equally arbitrary statistical one, and can produce weights that are difficult to explain

to the board member reading the dashboard, or weights that shift meaningfully from one reporting period to the next simply because the underlying correlation structure across six LEAs moved, with no change in anyone's actual judgment about what matters. Budget-allocation and expert-elicitation weighting make the underlying value judgment visible by asking a defined group of people to distribute a fixed number of points across indicators and defend the distribution, trading the appearance of statistical neutrality for a record of who decided what and why.

BHPI's working plan uses expert elicitation, conducted through the same structured consultation format as the Activity 2.3 crosswalk sessions, with the weighting exercise run separately by domain across all six LEA teams during Activity 2.6. Running the exercise separately by domain, rather than asking each team to weight all four domains' indicators in one sitting, keeps the elicitation task cognitively manageable and keeps a team's judgment about, say, outcome indicators from bleeding into their judgment about demographic indicators simply because both were asked back to back. That keeps the weighting judgment traceable to named participants and a documented process rather than to an unexplained statistical artifact, and it means the published methodology can say who set the weights and how, which equal weighting and PCA weighting cannot say. Mandatory sensitivity analysis on the resulting weights is not optional under the OECD/JRC framework; a composite whose ranking of districts flips under a modest reweighting has not demonstrated anything robust (OECD ET AL. 2008), and the Activity 2.6 deliverable will include a Monte Carlo perturbation of both the weights and the normalization choice, reporting how much a district's BHPI score and rank move under plausible alternative specifications.

Aggregation And The Compensability Question

Once indicators are normalized and weighted, they still have to be combined, and how they are combined decides whether strength in one place can cover for weakness in another. Linear (additive) aggregation is fully compensatory: a high score on one indicator can offset a low score on another indicator at whatever exchange rate the weights imply. Geometric aggregation is partially non-compensatory: because it multiplies rather than sums normalized components, a very low score on any one component drags the whole product down disproportionately, so no amount of strength elsewhere fully covers for it (OECD ET AL. 2008), and the OECD/JRC framework is explicit that this is not a stylistic preference but a policy-relevant fork with genuinely different implications for a floor-sensitive construct.

That distinction is not academic for a behavioral health index specifically. A linear, fully compensatory BHPI would let a district average away a real gap, for instance a crisis-response or service-completion shortfall, against strong attendance or participation numbers elsewhere in the same domain, producing a headline score that looks fine while masking exactly the kind of gap the Commission most needs to see.

SECTION 08 · INDEX METHODOLOGY · CONTINUED

Aggregation And The Compensability Question

BHPI v0.9 therefore leans toward geometric aggregation within domains where a floor genuinely matters (service utilization and outcomes/impact), reserving linear aggregation for domains where indicators are closer to genuinely substitutable (demographics/characteristics), on the reasoning that a demographic profile's component figures describe who a district serves rather than how well it is performing, and substitutability among descriptive figures does not carry the same stakes as substitutability among performance figures. The final aggregation rule per domain is one of the specific items the Activity 2.6 structured review will confirm or revise with all six LEA teams before publication, and the worked example further down this page shows concretely why the choice cannot be settled by intuition alone: geometric aggregation is undefined for a negative normalized value, which means adopting it for a domain also commits BHPI to a rescaling step the Activity 2.6 review specifies.

From Score To Color: Performance Banding

A continuous index score is precise, but precision is not the same thing as actionable. For a board member deciding where to direct attention, a single decimal value is harder to act on than a small number of clearly defined performance bands. BHPI proposes adapting the design California school districts already use for the state's own accountability reporting: the California School Dashboard's five-by-five status-and-change grid, which crosses a percentile-based status level against a year-over-year change level to place each indicator into one of twenty-five color-coded cells, spanning five performance colors from blue at the top of the grid through green, yellow, and orange to red at the bottom, with cut points published in the accompanying technical tables

CALIFORNIA DEPARTMENT OF EDUCATION. 2025. Reusing that grid logic for the BHPI means every LEA reads behavioral health performance in the same visual grammar they already use for academic and climate indicators on the state Dashboard, rather than learning a second, unrelated color system, and it means a superintendent's cabinet does not need a new legend the first time BHPI appears next to their existing Dashboard tiles.

The status axis places a district's current BHPI (and each sub-index) relative to the six-LEA distribution for the current reporting period. The change axis places the year-over-year movement in that same score. A district can land, for example, in a band that reads "solid current performance, improving" or "below-median current performance, declining," a framing that separates the two questions a superintendent actually has, where do we stand and are we getting better, rather than collapsing both into one number. Cut points for both axes, and the exact banding table, are among the deliverables the Activity 2.6 review will finalize; this page publishes the resulting table once that review sets it. It is also worth being plain that CDE's own cut points were calibrated against the full population of California's roughly one thousand school districts, and a direct import of those specific numeric thresholds onto a six-LEA reference distribution would not be defensible; what BHPI borrows from the Dashboard is the grid's structure and color grammar, not its specific percentile cut scores, which have to be recalculated against BHPI's own six-LEA distribution.

The change axis raises a measurement question the OECD/JRC handbook does not fully resolve, because computing "improvement" is a harder problem than computing "status": a naive year-over-year subtraction treats every district's prior-year score as equally reliable, which is not true when reporting periods, cohort sizes, or administration timing differ across the six LEAs. The literature on relative-rating systems offers a more rigorous alternative worth flagging for the Activity 2.6 review: the paired-comparison rating framework originally developed for competitive ranking, which estimates relative skill from win-loss outcomes against opponents of known rating (**ELO 1978**), and refined into a dynamic setting that models a competitor's ability as a value that changes over time via a non-linear state-space process, fit with a computationally simple, non-iterative algorithm that explicitly represents how uncertain each rating estimate is rather than treating every observed value as equally precise (**GLICKMAN 1999**). Adapting that logic to a change-level calculation would mean weighting a district's year-over-year movement by how confidently each period's score was measured (a period built on a fuller quarter of data would carry more weight in the change calculation than a period built on a partial or newly onboarded quarter), rather than by simple subtraction, a refinement to consider for the change axis specifically, separate from and in addition to the four-domain composite construction this page otherwise describes. This is flagged as literature worth considering, not a construction decision BHPI has made; whether the added statistical complexity of a state-space change model is worth it for a six-LEA reporting population this size is itself a question the Activity 2.6 review should weigh against the simpler, more transparent alternative of a straightforward year-over-year difference reported alongside its own confidence interval.

Composite Reliability: Omega, Not Alpha

A composite index should report how reliably its components hang together, the same way any measurement instrument reports internal consistency. The conventional choice, Cronbach's alpha, assumes the components are tau-equivalent, that each indicator measures the underlying construct with the same precision and the same units of "true score." That assumption does not hold for a domain sub-index that mixes a completion count, a fidelity rating, and a service-utilization rate, each with a different relationship to the construct it is meant to represent, and when the tau-equivalence assumption is violated, alpha is a known-biased estimate of the sub-index's actual reliability rather than merely a slightly-off one. Coefficient omega, derived from a factor-analytic rather than tau-equivalent measurement model, remains a defensible reliability estimate precisely under the heterogeneous-component condition where alpha is not (**MCDONALD 1999**). BHPI will report omega for each domain sub-index and for the composite BHPI itself, alongside the correlational coherence check described above (the two are related but distinct checks: coherence asks whether a domain's indicators belong together at all, omega asks how reliably they measure whatever they belong together as, once that grouping is settled), rather than defaulting to alpha because it is the more familiar figure.

SECTION 08 · INDEX METHODOLOGY · CONTINUED

Composite Reliability: Omega, Not Alpha

It is worth separating this composite-level reliability question from a different reliability figure that will come up elsewhere on this site: the scoring-model agreement statistics reported on the current Model Card, which describe how consistently Impacter Pathway's ordinal scoring pipeline (Model Suite v6.0) agrees with trained human raters on an individual student response, an average quadratic-weighted kappa of 0.918 across the eight competency attributes, ranging by attribute from 0.893 to 0.952. That figure answers "does the machine learning model score a single response the way a human would." Composite reliability, by contrast, answers "do the indicators inside a domain sub-index cohere as a group." A high omega does not certify individual scoring accuracy, and a high scoring-model kappa does not certify that a domain's indicators belong together; the two are complementary but not substitutes for each other. They should not be conflated, and neither figure should be cited as if it stood in for the other in a board presentation where only one number fits on a slide.

Uncertainty Quantification And The Mandatory Sensitivity Check

Every choice documented above, which normalization, which weights, which aggregation rule, is a specific decision among several defensible alternatives, and the OECD/JRC framework treats reporting a single point estimate without testing its sensitivity to those choices as incomplete methodology, not merely good practice (OECD ET AL. 2008). The specific statistical technique behind that requirement, rather than sensitivity analysis as a vague gesture toward "checking robustness," is the pairing of uncertainty analysis with variance-based sensitivity analysis that a JRC team formalized for composite-indicator construction and demonstrated on international benchmarking indices, using a generalization of Sobol's method to apportion how much of a composite's ranking uncertainty traces to one specific construction choice, such as the weighting scheme, versus another, such as the missing-data imputation rule (SAISANA ET AL. 2005). That distinction matters for what the Activity 2.6 deliverable should actually report: not one undifferentiated robustness number, but an apportioned answer to the question "if a district's BHPI score or rank is unstable, which specific methodological choice is driving that instability," so a district or a reviewer knows which construction decision to scrutinize rather than being told only that the score is somewhat uncertain overall.

The Activity 2.6 deliverable accompanying this page will include a documented sensitivity analysis along those lines: a Monte Carlo procedure that perturbs the normalization method, the weight set, and the aggregation rule across a defined range of plausible alternatives, and reports how much each district's BHPI score and relative rank move as a result, decomposed by which perturbed input is responsible for how much of that movement. A district whose ranking is stable across that perturbation range has a defensible score; a district whose ranking flips under small, reasonable changes to the construction method has a score that carries real caveats

rather than false precision, and the apportioned sensitivity result should say specifically whether that instability traces to the weight set, the normalization choice, or the aggregation rule, since the appropriate fix differs depending on which one it is. Every BHPI figure referenced elsewhere on this site carries the v0.9 version marker the banner above states, and the sensitivity analysis is reported alongside it.

Protecting Students: Suppression And The No-Student-Level-Data Rule

The public-facing BHPM dashboard reports aggregate, district- and school-level figures only; it does not and will not publish student-level data, and every domain indicator is defined at a level of aggregation designed to prevent indirect re-identification of an individual student through a small reporting cell. Federal guidance for statewide longitudinal data systems documents that most states set a minimum reporting-group size between five and thirty students for exactly this reason, with ten the most common threshold, and recommends that states adopt and publicly state a specific rule rather than leaving suppression to case-by-case judgment (SEASTROM 2010). That same guidance also flags that threshold suppression alone can still allow disclosure by subtraction across related cells unless paired with complementary suppression: if a school's total enrollment and every demographic subgroup but one are published, a reader can back out the suppressed subgroup's value by subtracting the published ones from the total, even though that one cell never appeared on the page in numeric form. Complementary suppression closes that gap by suppressing at least one additional cell in the same related set whenever a single cell falls below threshold, so no combination of published figures lets a reader reconstruct the number that was withheld, which is a design detail the BHPI dashboard's suppression logic needs to account for and not just a single minimum-n check.

The working default for BHPI is a minimum reporting-group size of ten students per cell, following the modal practice the federal guidance documents, applied together with complementary suppression across related demographic and outcome breakdowns. That default runs through SCCOE data governance before it becomes the production rule, and this page carries the rule once that approval is recorded. Until then, any cell falling below the working threshold displays as suppressed rather than as an estimated or rounded figure, and no component of the BHPI construction pipeline retains or exposes student-level records outside the platform's existing FERPA- and COPPA-governed data environment. A suppressed cell and a not-yet-reported cell should read differently to anyone looking at the dashboard, since one reflects a privacy protection working as designed and the other reflects a data-collection gap that the platform should eventually close; collapsing the two into a single blank or dash would erase a distinction the reader actually needs.

SECTION 08 · INDEX METHODOLOGY

Keeping The Method Honest: Versioning And Public Documentation

The final step of the OECD/JRC sequence is presentation built for the audience that will actually use it, and a version-controlled, publicly accessible methodology page is how BHPI keeps that commitment durable rather than one-time (OECD ET AL. 2008). This page carries an explicit version marker (v0.9, settled by the Activity 2.6 review) precisely so that a change to normalization, weighting, aggregation, or banding produces a new version number and a visible changelog rather than a silent edit to a score's definition. The same discipline applies to the underlying scoring pipeline referenced above: model versions evolve over time and are versioned and audited for traceability and reproducible scoring, and any figure quoted from a specific model version on this site is tied to that version rather than presented as a permanent constant. A reader returning to this page in a year should be able to tell exactly what changed, when, and under whose review, the same standard the California Department of Education holds itself to by archiving each year's Dashboard technical guide separately rather than overwriting the prior version (CALIFORNIA DEPARTMENT OF EDUCATION. 2025). Concretely, that means a v1.0 release of this page at the close of Activity 2.6 will retain v0.9 as an archived prior version rather than replacing it outright, so that any BHPI figure already reported to the Commission under v0.9 stays traceable to the exact methodology that produced it, even after the production methodology has moved on.

Activity 2.4 · The Real BHPI, As Already Reviewed With The Commission

Everything above works through the general OECD/JRC composite-construction sequence in the abstract: the families of options BHPI could draw on for normalization, weighting, and aggregation. Activity 2.4's actual job is narrower and more concrete: define the specific composite BHPI is reporting. That composite already exists in working form, was presented to Pilar and SCCOE alongside the "v2" dashboard and educator-portal prototypes on 2026-08-07, and is reused here exactly rather than re-derived from the general framework above. Unlike the aggregation discussion two sections up, BHPI's realized construction is a deliberate, fully linear, compensatory weighted sum, not a geometric or otherwise non-compensatory aggregation, a specific choice for this specific composite, not a contradiction of the general aggregation-family literature reviewed above, which still applies to how BHPM evaluates aggregation choices domain by domain elsewhere in the framework.

BHPI is Domain 4's flagship outcome metric (the number Outcomes & Impact reports), not a fifth index competing with the Commission's four-domain taxonomy. Domains 1 through 3 (services provided, student demographics and characteristics, service utilization) remain the input and context reporting categories built out under Activity 2.3; Domain 4 is where this composite gets reported.

Beneath the Voice dimension sits the per-student Wellness Index, scored across six BHPM-specific competencies: Self-Insight, Emotional Resilience, Relational Awareness, Conflict Resolution, Effective Help-Seeking, Reflective Growth. These six competencies were developed specifically for BHPM, co-developed with the San Diego County Office of Education, a distinct, BHPM-specific set, not the general Impacter Pathway 8 competency model used elsewhere on the platform.

The bands above translate a continuous BHPI score into the same kind of actionable read the general banding discussion above calls for, but they are BHPM's own four-level scale (Leading, Establishing, Developing, Emerging) set against fixed cut points rather than the California Dashboard's percentile-based five-by-five grid described earlier on this page. Both are strengths-based, plain-language framings; which one the production dashboard ships with, or whether the two get reconciled into one grammar, is itself one of the specific questions Activity 2.6's structured review with all six LEA teams will settle, consistent with this page's v0.9 version marker. Publishing the real formula here does not change that version: weighting, banding, and the domains' relationship to each other still finalize through Activity 2.6, and this composite is presented as v0.9, under Commission and partner review.

The Per-Student Rubric Beneath Voice: Six Competencies, Four Levels

The "layer beneath Voice" panel above names the six BHPM-specific Wellness Index competencies (Self-Insight, Emotional Resilience, Relational Awareness, Conflict Resolution, Effective Help-Seeking, Reflective Growth) as a badge list. What follows is the actual rubric behind those six names: the four-level Foundational / Developing / Competent / Advanced scoring guide a rater (human or machine-learning-scored) applies to a single student's single spoken reflection on a single competency.

This is a different scale from the BHPI performance bands shown above, even though both happen to have four levels, and the two should not be conflated. The rubric below scores one student, on one competency, from one piece of authentic student speech. It lives entirely inside the Voice component, well below the level BHPI itself reports at. The Leading / Establishing / Developing / Emerging bands above score an LEA's (or the composite's) aggregate BHPI value on a 0-1 axis, at the site level, once every student's scores across every dimension have already rolled all the way up. Four levels each is a coincidence of scale design, not a shared instrument: a student rated Competent on Self-Insight is not "the same" as a district banded Establishing on BHPI, and this page's two four-level scales should be read as two separate rubrics measuring two separate things, not one rubric reported twice.

SECTION 09 · STAKEHOLDER ENGAGEMENT

Stakeholder Engagement And The Local-Framework Crosswalk

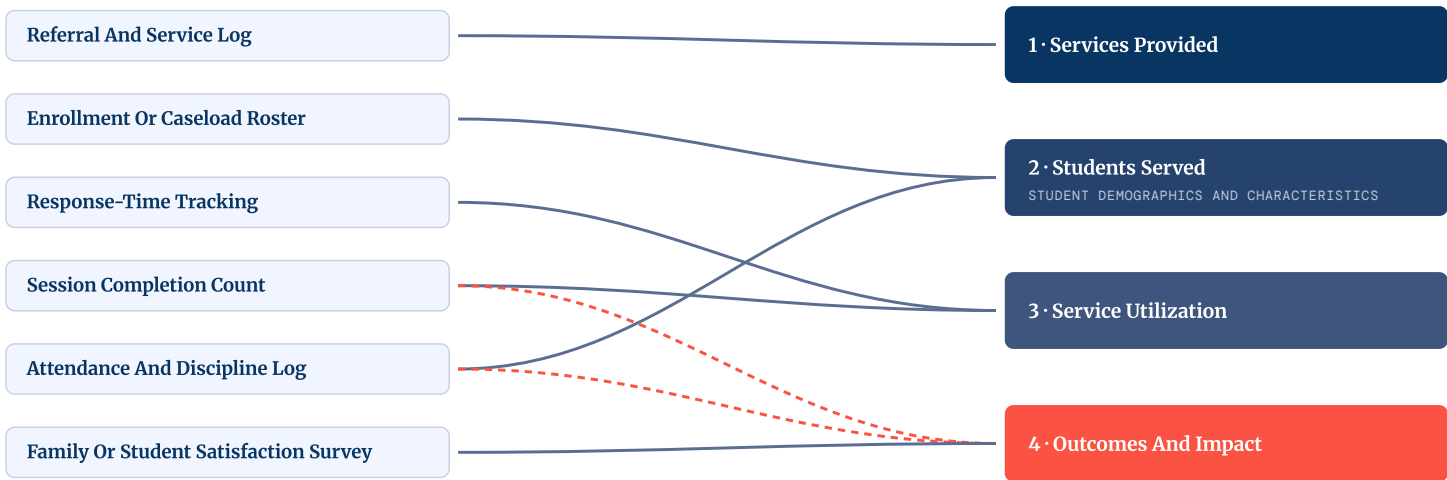
SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.2: structured sessions led by Deveau Burr Group with LEA behavioral health leads, counselors, site administrators, county behavioral health staff and BHSSA grantees, crosswalking each LEA's existing local framework onto the Commission's four reporting domains.

Activity window 07/01/26–07/30/26 · led by Deveau Burr Group (DBG), with Santa Clara County Office of Education clinical staff and Breaking Barriers California support, Impacter Pathway in a supporting role · session cadence and count carried in the Deveau Burr Group session log.

How A District's Own Terms Land In The Commission's Four Domains

The crosswalk pattern. A solid line is where a local measure maps; a dashed line is a measure that straddles two domains and stays mapped to both.



WHAT THE SECTION COVERS

- Who's In The Room
- Two Things These Sessions Produce
- Where The Session Record Lives

The full section, with every source beside the prose: bhpm.care/deliverables/d2/stakeholder-engagement

KEY TAKEAWAY · THE RECORD'S OWN SUMMARY

An indicator earns its place in the S4 candidate indicator table, and in the Index Methodology page's domain sub-index structure, because a crosswalk consultation traced it back to something an LEA already reports, not because it read well in the literature. The Index Methodology page's weighting exercise runs on that same structured consultation format: DBG-facilitated, per-domain, participants named and the process documented, which is what makes the weighting exercise defensible. Every indicator in the S4 table traces back to an Activity 2.2 session. And a measure that genuinely straddles two Commission domains, like the attendance and discipline log above, stays mapped to both rather than getting resolved into a single domain for the sake of a cleaner table.

SECTION 09 · STAKEHOLDER ENGAGEMENT · CONTINUED

Stakeholder Engagement And The Local-Framework Crosswalk

Activity status

Activity window 07/01/26–07/30/26 · led by Deveau Burr Group (DBG), with Santa Clara County Office of Education clinical staff and Breaking Barriers California support, Impacter Pathway in a supporting role · session cadence and count carried in the Deveau Burr Group session log.

A literature review can tell BHPM what other systems have measured. It cannot tell BHPM what the six people running behavioral health support at a specific school actually think is worth counting, or where their district's own paperwork already tracks something the Commission's framework has a different name for. That is what Activity 2.2 is for. Where Activity 2.1 built the evidence base from the outside, Activity 2.2 goes into the room with the people doing the work and asks them directly, then does the unglamorous translation work of lining up what they already track against the four categories the Commission needs reported.

Who's In The Room

DBG convenes structured sessions with four groups, drawing on SCCOE clinical staff and Breaking Barriers California to bring both a county behavioral health perspective and a community-partner perspective into the same conversation rather than collecting them separately:

County behavioral health department staff: the clinical and administrative side of how BHSSA-funded services actually get delivered and billed.

LEA behavioral health leads: the people accountable for a district's program end to end.

Counselors and site administrators: the school-level staff closest to a student's actual support pathway, and the first to know when a measure everyone agreed on upstream doesn't hold up in practice.

BHSSA grantees: the broader set of grant-funded partners whose work the indicator set also has to account for.

Two Things These Sessions Produce

Meaningful process and outcome measures, named by the people who'd have to report them. A measure a district's own staff can explain, defend, and actually collect without a new system is worth more than one that only looks clean in a literature review. These sessions are where a candidate indicator gets tested against that standard before it earns a place in the Activity 2.3 crosswalk table.

Crosswalk consultations that map each LEA's existing local framework onto the Commission's four reporting domains. Every district already tracks something (a referral log, a caseload roster, a completion count) in its own language, built for its own workflow, long before BHPM asked it to report against services provided; student demographics and characteristics, referred to on this site as Students Served; service utilization; and outcomes and impact. The consultation's job is not to replace that local language. It is to find where each local term already lands inside the Commission's four categories, and to be honest when one local measure genuinely straddles two of them rather than forcing a single tidy answer.

Crosswalk pattern

How a district's own local terms map onto the Commission's four domains, including the terms that land in two at once. It is drawn as a diagram in the earlier working design.

[Open crosswalk pattern in the archive](#)

That splitting is not an edge case to engineer around. An attendance and discipline log is a good example: it describes who is affected, which belongs in Students Served, and it also functions as an outcome covariate, which belongs in Outcomes & Impact, the same dual role S4's discussion of contextual indicators names for exactly this kind of measure. A crosswalk consultation that forced every local term into exactly one domain would be tidier to publish and less honest about what the underlying data actually does.

Where The Session Record Lives

The crosswalk sessions and the measures-identification sessions ran within the 07/01/26–07/30/26 window per the Implementation Workplan Crosswalk. Deveau Burr Group holds the session log, which carries the count, the cadence and the per-LEA breakdown, together with the local framework terms each district's staff surfaced. The diagram above illustrates the mapping pattern these consultations use rather than any one district's own language.

SECTION 10 · THE METRICS BRIEF

The Metrics Brief · Version 2a

SCOPE · IMPLEMENTATION WORKPLAN CROSSWALK

Activity 2.5: synthesize Activities 2.1 through 2.4 into the Metrics Brief: the proposed indicator set by domain, the index, and the privacy, security and data governance assessment, carried into LEA structured review under Activity 2.6.

The Proposed Indicator Set, By Domain

INDICATOR	FAMILY	COMMISSION DOMAIN	FEASIBILITY, AS THE BRIEF STATES IT
Screening Administration Completion Rate	Process	Services Provided	HIGH captured automatically by the platform; no manual entry.
Implementation Fidelity Score (TFI-Aligned)	Process	Services Provided	MODERATE requires a trained local scorer and a fixed administration cadence.
Staff And Educator Training Participation Rate	Process	Services Provided	HIGH logged through existing TTA delivery tracking.
Data-Review Cycle Completion	Process	Services Provided	HIGH once a site's review calendar is established in D1.
Screened Population Profile (Grade Band, Language, Program Status)	Contextual	Students Served	HIGH fields already reside in roster feeds.
Baseline Tier And Need-Level Distribution	Contextual	Students Served	HIGH
English Learner And IEP/504 Status Cross-Tab On Screening Participation	Contextual	Students Served	MODERATE depends on the site's data-sharing agreement covering those fields.
Referral-To-Service Rate	Utilization	Service Utilization	MODERATE depends on district-provider data-sharing agreements.
Billable Screening Encounters (CPT 96127)	Utilization	Service Utilization	HIGH where a site's billing infrastructure already supports the code; low where none exists yet.
School-Linked Provider Network Utilization	Utilization	Service Utilization	DEPENDS Depends on the district's CYBHI cohort enrollment status.
Tier Movement Over Time	Outcome	Outcomes And Impact	HIGH once two screening administrations exist.
Exclusionary Discipline Rate (Suspensions And Expulsions)	Outcome Covariate	Outcomes And Impact	HIGH already collected for state reporting.
Domain-Level Growth On Measured Competencies	Outcome	Outcomes And Impact	HIGH computed automatically by the scoring pipeline.

WHAT THE SECTION COVERS

- How The Indicator Set Was Identified
- The Proposed Indicator Set, By Domain
- Aligning The Indicator Set: Statutory Goals, CYBHI Billing, And LEA Frameworks
- The Behavioral Health Performance Index
- Indicator Count Per Domain
- Privacy, Security & Data Governance

The full section, with every source beside the prose: bhpmmcare.com/deliverables/d2/brief

SECTION 10 · THE METRICS BRIEF · CONTINUED

Privacy, Security And Data Governance, And What Comes Next

PRIVACY, SECURITY AND DATA GOVERNANCE

The Commission requires the brief to assess the privacy and security implications of the data collected (its sources, its flows, its collection methods) and to establish data governance processes giving explicit operational guidance to everyone handling it. That section is written in full in the earlier working design and carries forward here. Open the data governance section in the archive

NEXT STEPS

Activity 2.6 runs LEA structured review with all six LEA teams. That review finalizes the open cadence and data-source questions flagged in the table above and confirms BHPI's weighting, banding, and sensitivity analysis. The result is Deliverable 2b, the final Brief submitted to the Commission.

VERSION

Deliverable 2a · for LEA structured review · Activity 2.5, 08/01/26–08/29/26 · becomes Deliverable 2b through the Activity 2.6 review and Commission revision.

2a

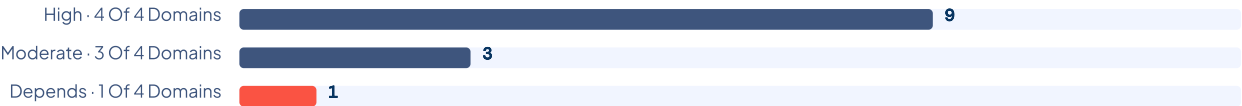
This Brief, For LEA Structured Review

2b

The Final Brief, After Activity 2.6

Feasibility, Counted Across The Thirteen

How many indicators the brief rates at each level of feasibility. An indicator rated High can be produced today; Moderate needs a scorer, a cadence or an agreement; Depends waits on billing or data-sharing infrastructure.



4 Services Provided	3 Students Served	3 Service Utilization	3 Outcomes And Impact	13 Candidate Indicators In Total
------------------------	----------------------	--------------------------	--------------------------	-------------------------------------

The Metrics Brief · Version 2a

Version

Deliverable 2a · for LEA structured review · Activity 2.5, 08/01/26–08/29/26 · becomes Deliverable 2b through the Activity 2.6 review and Commission revision.

BHPM answers the question it exists to answer for the Commission: what does a school behavioral health performance management system count, and how is it reported. It proposes an indicator set, organizes that set under the Commission's four reporting domains, and names the one composite score BHPM reports as its flagship outcome measure: the Behavioral Health Performance Index.

It synthesizes four completed activities. The S1–S7 literature review (Activity 2.1), DBG's stakeholder engagement sessions with the six LEAs and county behavioral health staff (Activity 2.2), the domain–alignment work sorting indicators into the Commission's taxonomy (Activity 2.3), and the BHPI index methodology (Activity 2.4) all come

together here in one record. Every indicator below carries a definition, a data source, a cadence, and a citation back to the S1–S7 evidence base; click any row for the full detail and its sources. Where a cadence or a data-sharing dependency governs an indicator, the table names it.

How The Indicator Set Was Identified

The Commission's later Final Report deliverable will ask this Brief a direct question: not just which indicators BHPM proposes, but a description of how they were identified. That is not a rhetorical requirement this page can satisfy with a sentence. It is five linked activities, each with its own page on this site, synthesized here rather than replaced by this synthesis.

How the indicators were identified

The five linked activities, in order, that produced the candidate indicator set. It is drawn as a diagram in the earlier working design.

[Open how the indicators were identified in the archive](#)

Literature review (Activity 2.1) built the evidence base from published research first, before any indicator was proposed to a single stakeholder. Stakeholder engagement (Activity 2.2) then tested that candidate set against the people who would actually report it, and ran the crosswalk consultations mapping each LEA's own local framework onto the Commission's four domains, including the split-mapping pattern the [Stakeholder Engagement page](#)'s crosswalk diagram documents for a measure like an attendance and discipline log, which describes who is affected and also functions as an outcome covariate at once. Domain alignment (Activity 2.3) sorted the resulting set into the Commission's four reporting domains. The Behavioral Health Performance Index (Activity 2.4) defined the one composite score built from that aligned set. This page, Activity 2.5, is

where all four synthesize into one record, and Activity 2.6 is where SCCOE runs LEA structured review of it before it becomes Deliverable 2b.

The Proposed Indicator Set, By Domain

Thirteen candidate indicators, the same working set the Activity 2.3 domain–alignment diagram sorts, read here through the Brief's own questions: what is it, where does the number come from, how often does it refresh, and can BHPM actually produce it today. "Students Served" below is the working shorthand for the crosswalk's formal domain name, student demographics and characteristics.

Services Provided · 4 Indicators

INDICATOR	DEFINITION	DATA SOURCE	CADENCE	FEASIBILITY
Screening administration completion rate	Share of enrolled, eligible students who completed a behavioral health screening administration.	Platform-logged automatically.	Per reporting window, per school.	High: captured automatically by the platform; no manual entry.
Implementation fidelity score (TFI-aligned)	A site's implementation fidelity, scored against the Tiered Fidelity Inventory across Tier 1–3 practices.	Trained site coordinator scoring, TFI or TFI-aligned local check.	Periodic, on the Tiered Fidelity Inventory administration cadence the site sets.	Moderate: requires a trained local scorer and a fixed administration cadence.
Staff and educator training participation rate	Share of the site staff roster who completed BHPM onboarding and technical-assistance modules.	Existing TTA delivery tracking.	On the training and technical assistance delivery cadence.	High: logged through existing TTA delivery tracking.
Data-review cycle completion	Whether a site is holding its scheduled behavioral health data-review meetings.	Meeting count against the site's D1 implementation plan calendar.	Set by each site's own Deliverable 1 implementation plan calendar.	High once a site's review calendar is established in D1.

SECTION 10 · THE METRICS BRIEF · CONTINUED

The Proposed Indicator Set, By Domain

Student Demographics And Characteristics · 3 Indicators

INDICATOR	DEFINITION	DATA SOURCE	CADENCE	FEASIBILITY
Screened population profile (grade band, language, program status)	Demographic composition (grade band, language, program status) of the students who were screened.	Enrollment/roster feeds, cross-tabulated against screening completion.	Aligned to each screening administration window.	High: fields already reside in roster feeds.
Baseline tier and need-level distribution	Count of students at each MTSS tier as of first screening administration, by school.	Screening administration records.	One-time baseline, at first screening administration.	High.
English learner and IEP/504 status cross-tab on screening participation	Screening participation rate broken out by English learner and IEP/504 status.	Roster fields cross-tabulated with screening records. Depends on the site's data-sharing agreement covering those fields.	Aligned to the screening cadence.	Moderate: depends on the site's data-sharing agreement covering those fields.

Service Utilization · 3 Indicators

INDICATOR	DEFINITION	DATA SOURCE	CADENCE	FEASIBILITY
Referral-to-service rate	Share of screened students referred for behavioral health service, tracked to service entry where data-sharing allows.	Referral tracking, matched to provider records where a data-sharing agreement exists.	Set by the district-provider data-sharing agreements.	Moderate: depends on district-provider data-sharing agreements.
Billable screening encounters (CPT 96127)	Count of standardized-instrument screening encounters coded and documented for reimbursement.	Site billing system, CPT 96127.	Per billing cycle.	High where a site's billing infrastructure already supports the code; low where none exists yet.
School-linked provider network utilization	Count and rate of behavioral health encounters reimbursed under the DHCS statewide fee schedule.	DHCS CYBHI Fee Schedule Program billing data, at CYBHI-enrolled sites.	Set by the district's CYBHI cohort enrollment status.	Depends on the district's CYBHI cohort enrollment status.

Outcomes And Impact · 3 Indicators

INDICATOR	DEFINITION	DATA SOURCE	CADENCE	FEASIBILITY
Tier movement over time	Change in a student's MTSS tier placement between baseline and a later screening, aggregated to school level.	Screening administration records, baseline vs. follow-up, with small-cell suppression applied.	Each subsequent screening window.	High once two screening administrations exist.
Exclusionary discipline rate (suspensions and expulsions)	Suspension and expulsion rate, cross-referenced against BHPM tier and fidelity data.	District SIS discipline records, already collected for state reporting.	Aligned to the district's existing state discipline-reporting cycle (not independently specified here).	High: already collected for state reporting.
Domain-level growth on measured competencies	Within-student change in scored competency-attribute levels across successive administrations.	Impacter Pathway scoring pipeline output.	Recalculated at each administration.	High: computed automatically by the scoring pipeline.

Two domains carry most of the feasibility risk worth naming up front. Service Utilization depends the most on infrastructure outside BHPM's own platform: a district's CYBHI cohort enrollment status, its billing system's support for CPT 96127, or a provider data-sharing agreement. So its indicators appear as reporting categories with a feasibility flag, not as promised figures every site will have from day one. Students Served carries the indicator set's disaggregation obligation: the one independent randomized trial in this literature found its real effects concentrated entirely inside a highest-need subgroup, invisible in a school-wide average, which is why demographic and need-tier breakdowns are written into the indicator set itself rather than added later as an equity appendix.

Aligning The Indicator Set: Statutory Goals, CYBHI Billing, And LEA Frameworks

Naming thirteen indicators is not, on its own, what Activity 2.5 is required to deliver. Each one also has to line up with three separate things at once: the reason a Commission exists to hold this deliverable accountable, the billing infrastructure some sites already run through, and the language an individual LEA staffer actually uses. Those three don't collapse into a single column on the table above without losing something true about at least one of them, so this Brief keeps them separate here instead.

Three alignment frames

Statutory goals, CYBHI billing categories and each LEA's own framework language, held side by side. It is drawn as a diagram in the earlier working design.

[Open three alignment frames in the archive](#)

SECTION 10 · THE METRICS BRIEF

The Behavioral Health Performance Index

Domain 4, Outcomes and Impact, carries BHPM's flagship score. The full construction rationale, covering normalization, weighting, aggregation, banding, and disclosure rules under the OECD/JRC composite-construction framework, lives on the Index Methodology page. The working formula, performance bands, and the Wellness Index layer beneath it appear here because thirteen indicators mean little without the one composite score built from them.

Beneath the Voice dimension sits the per-student Wellness Index, scored across six BHPM-specific competencies: Self-Insight, Emotional Resilience, Relational Awareness, Conflict Resolution, Effective Help-Seeking, Reflective Growth. These six competencies were developed specifically for BHPM, co-developed with the San Diego County Office of Education, a distinct, BHPM-specific set, not the general Impacter Pathway 8 competency model used elsewhere on the platform.

Indicator Count Per Domain

The count is uneven by design, not by omission. Services Provided carries the most indicators of the four right now because it's the

domain BHPM's own platform already logs directly: screening, fidelity, and training activity with no manual entry. Students Served, Service Utilization, and Outcomes and Impact sit tied at three apiece, and Service Utilization is the domain most worth watching: its three indicators carry the heaviest feasibility caveats on the table above, since each one waits on billing or data-sharing infrastructure some sites don't have yet. That gap is a fair target for the Activity 2.6 review to close as more sites clear their CYBHI and billing readiness bar.

Privacy, Security & Data Governance

This is a named Commission requirement for Activity 2.5, not a courtesy inclusion: this section assesses the privacy and security implications of the data BHPM proposes to collect, covering its sources, its flows and its collection methods, and establishes data governance processes that give explicit operational guidance to everyone handling it. Nothing below is a new policy invented for this page; it is the architecture this site and the RFP's own DI recruitment and participation-agreement activities already establish, organized under the three categories the Commission's own requirement names.

Privacy, security & data governance

The Commission requires the brief to assess the privacy and security implications of the data collected (its sources, its flows, its collection methods) and to establish data governance processes giving explicit operational guidance to everyone handling it. That section is written in full in the earlier working design and carries forward here.

[Open the data governance section in the archive](#)

Next Steps

Activity 2.6 runs LEA structured review with all six LEA teams. That review finalizes the open cadence and data-source questions flagged in the table above and confirms BHPM's weighting, banding, and sensitivity analysis. The result is Deliverable 2b, the final Brief submitted to the Commission.

Prepared for the California Behavioral Health Services Oversight & Accountability Commission (BHSOAC) under the Behavioral Health Student Services Act (BHSSA), Contract No. 25BHSOAC019. Santa

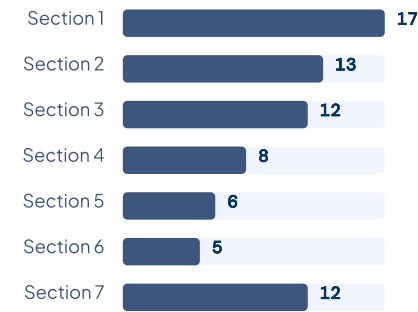
Clara County Office of Education (administering) · Impacter Pathway · Stanford Medicine · Applied Survey Research · Breaking Barriers California.

SECTION 11 · BIBLIOGRAPHY

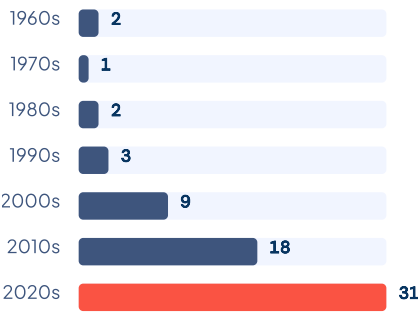
The Evidence Base

Every source cited across the seven literature-review sections, grouped by the section that cites it, with the provenance and verification date carried from the source registry.

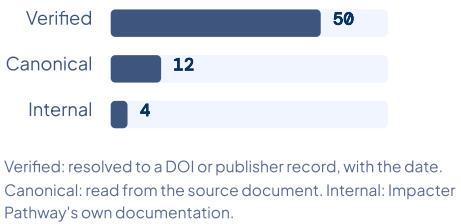
SOURCES CITED, BY SECTION



SOURCES, BY DECADE PUBLISHED



PROVENANCE



Index Of Sources

Every entry in the registry, with the literature sections that cite it. The annotated entries follow, section by section.

American Educational Research Association et al., 2014 Section 3	Goodman, 1997 Section 1	McIntosh et al., 2023 Section 4
American Medical Association., 2026 Section 4	Granger, 1969 Section 7	Ober et al., 2026 Section 2, Section 7
Anderson et al., 2019 Section 1	He et al., 2021 Section 2, Section 3	OECD et al., 2008 Section 6
Arnold et al., 2007 Section 7	He et al., 2021 Section 2	Palmer et al., 2025 Section 1
Baffsky et al., 2023 Section 5	Hirshberg et al., 2022 Section 7	Pennebaker et al., 1986 Section 7
Basting et al., 2022 Section 1	Horner et al., 2010 Section 4	Posner et al., 2011 Section 1
Bauer et al., 2020 Section 2, Section 3	IMPACTER Education Corp., 2026 Section 2, Section 3, Section 7	Proctor et al., 2011 Section 5
Bryk et al., 2015 Section 5	IMPACTER Education Corp., 2026 Section 3, Section 7	Saab et al., 2009 Section 7
Cal. Welf. & Inst. Code § 5961.4 et al., 2024 Section 4	IMPACTER Education Corp., 2026 Section 7	Saisana et al., 2005 Section 6, index-methodology
California Department of Education., 2025 Section 6	IMPACTER Education Corp., 2026 Section 7	Sanh et al., 2019 Section 2
California Department of Health Care Services., 2026 Section 4	Kilgus et al., 2013 Section 1	Scott et al., 2019 Section 4
Cao et al., 2020 Section 2, Section 3, Section 7	Kilgus et al., 2016 Section 1	Seastrom, 2010 Section 6
Cohen, 1960 Section 3	Kilgus et al., 2018 Section 1	Siceloff et al., 2017 Section 1
Damschroder et al., 2009 Section 5	Krippendorff, 2004 Section 3	Soneson et al., 2020 Section 1
Danielson et al., 2023 Section 1	Kroenke et al., 2001 Section 1	Spitzer et al., 2006 Section 1
Deming, 1986 Section 5	Lane et al., 2019 Section 1	Substance Abuse and Mental Health Services Administration., 2026 Section 4
Duckworth et al., 2015 Section 1	Lopez et al., 2021 Section 2	Teferra et al., 2024 Section 2
Eklund et al., 2022 Section 1	Loukina et al., 2019 Section 3	Turpin et al., 2023 Section 2, Section 7
Elo, 1978 Index Methodology	Low et al., 2020 Section 2	WestEd., 2026 Section 1
Fisher et al., 2026 Section 2	McCaffrey et al., 2022 Section 3	Williamson et al., 2012 Section 3
Fixsen et al., 2005 Section 5	McDonald, 1999 Section 6	Wood et al., 2021 Section 3
Glickman, 1999 Index Methodology	McIntosh et al., 2017 Section 4	Zapata-Rivera et al., 2022 Section 2

SECTION 11 · BIBLIOGRAPHY · CONTINUED

The Annotated Bibliography

Every source cited across the seven literature-review sections, grouped by the section that cites it, with the provenance and

verification date carried from the source registry. 66 entries in all.

Section 1 · The Screener Landscape · 17 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. <i>Journal of General Internal Medicine</i> , 16(9), 606–613. doi10.1046/j.1525-1497.2001.016009606.x	Anchors the self-report clinical-screener column of S1's instrument comparison against which school behavioral-health screeners (BASC-3 BESS, SAEBRs, SDQ) are benchmarked for format and psychometrics.	VERIFIED verified 2026-08-25
Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. <i>Archives of Internal Medicine</i> , 166(10), 1092–1097. doi10.1001/archinte.166.10.1092	Completes the clinical self-report screener pairing (with PHQ-9) that frames S1's discussion of adult-clinical-model screeners now adapted for school use.	VERIFIED verified 2026-08-25
Eklund, K., Kilgus, S. P., Willenbrink, J. B., Collins, B., Gill, N., Weist, M. M., Porter, J., Lewis, T. J., Mitchell, B., & Wills, H. (2022). Evidence of the internal structure and measurement invariance of the BASC-3 Behavioral and Emotional Screening System Teacher Form. <i>Journal of Psychoeducational Assessment</i> , 40(8), 936–949. doi10.1177/07342829221116807	Supports the BASC-3 BESS Teacher Form's structural-validity claim in S1's screener comparison table.	VERIFIED verified 2026-08-25
Basting, E. J., Naser, S. C., & Goncy, E. A. (2022). Assessing the factor structure and measurement invariance of the BASC-3 Behavioral and Emotional Screening System Student Form across race/ethnicity and gender. <i>Assessment for Effective Intervention</i> , 48(1), 43–51. doi10.1177/15345084221095440	Supports fairness/equity claims for the BASC-3 BESS Student Form self-report format cited in S1's screener table.	VERIFIED verified 2026-08-25
Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for elementary grades. <i>School Psychology Quarterly</i> , 28(3), 210–226. doi10.1037/spq0000024	Documents the origin and initial validation logic of SAEBRs, one of S1's core universal screener comparison instruments.	VERIFIED verified 2026-08-25
Kilgus, S. P., Eklund, K., von der Embse, N. P., Taylor, C. N., & Sims, W. A. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRs) Teacher Rating Scale and multiple gating procedure within elementary and middle school samples. <i>Journal of School Psychology</i> , 58, 21–39. doi10.1016/j.jsp.2016.07.001	Provides the psychometric-defensibility evidence base cited for SAEBRs in S1's screener comparison table.	VERIFIED verified 2026-08-25
Kilgus, S. P., Bonifay, W. E., von der Embse, N. P., Allen, A. N., & Eklund, K. (2018). Evidence for the interpretation of Social, Academic, and Emotional Behavior Risk Screener (SAEBRs) scores: An argument-based approach to screener validation. <i>Journal of School Psychology</i> , 68, 129–141. doi10.1016/j.jsp.2018.03.002	Represents the argument-based validation framework S1 cites as the maturity benchmark for screener score interpretation.	VERIFIED verified 2026-08-25
Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. <i>Journal of Child Psychology and Psychiatry</i> , 38(5), 581–586. doi10.1111/j.1469-7610.1997.tb01545.x	Represents the internationally dominant free-to-use multi-informant screener in S1's instrument comparison, contrasted against Impacter Pathway's language-based approach in S2.	VERIFIED verified 2026-08-25
Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., Currier, G. W., Melvin, G. A., Greenhill, L., Shen, S., & Mann, J. J. (2011). The Columbia-Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. <i>American Journal of Psychiatry</i> , 168(12), 1266–1277. doi10.1176/appi.ajp.2011.10111704	Anchors the individual-actionability/crisis-response end of S1's screener comparison spectrum, distinct from broader risk-screening tools like SAEBRs or SDQ.	VERIFIED verified 2026-08-25
Lane, K. L., Oakes, W. P., Cantwell, E. D., Common, E. A., Royer, D. J., Leko, M. M., & Allen, G. E. (2019). Predictive validity of Student Risk Screening Scale—Internalizing and Externalizing (SRSS-IE) scores in elementary schools. <i>Journal of Emotional and Behavioral Disorders</i> , 27(4), 221–234. doi10.1177/1063426618795443	Represents the free/low-cost universal screener category in S1's cost-and-time column of the instrument comparison table.	VERIFIED verified 2026-08-25
WestEd. (n.d.). The California School Climate, Health, and Learning Survey (CalSCHLS) System, including the California Healthy Kids Survey (CHKS). California Department of Education.	Serves as S1's in-state precedent for anonymized, aggregate-only universal screening, distinguishing it from individually-actionable clinical screeners in the same comparison table.	CANONICAL verified 2026-08-25
Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. <i>Educational Researcher</i> , 44(4), 237–251. doi10.3102/O013189X15584327	Grounds S1's self-report method-effects structural gap and directly motivates S2's argument for authentic-voice/language-based measurement as a complementary method.	VERIFIED verified 2026-08-25
Soneson, E., Howarth, E., Ford, T., Humphrey, A., Jones, P. B., Thompson Coon, J., Rogers, M., & Anderson, J. K. (2020). Feasibility of school-based identification of children and adolescents experiencing, or at-risk of developing, mental health difficulties: A systematic review. <i>Prevention Science</i> , 21, 581–603. doi10.1007/s11121-020-01095-6	Substantiates S1's acceptability/implementation-barriers structural gap, particularly the staff-buy-in dimension addressed later in S5's implementation-science section.	VERIFIED verified 2026-08-25
Anderson, J. K., Ford, T., Soneson, E., Thompson Coon, J., Humphrey, A., Rogers, M., Moore, D., Jones, P. B., Clarke, E., & Howarth, E. (2019). A systematic review of effectiveness and cost-effectiveness of school-based identification of children and young people at risk of, or currently experiencing mental health difficulties. <i>Psychological Medicine</i> , 49(1), 9–19. doi10.1017/S0033291718002490	Supports S1's framing that school screening's evidence base on downstream effectiveness, not just detection, is still developing, motivating S4/S5's emphasis on process-outcome linkage.	VERIFIED verified 2026-08-25
Palmer, C., Kane, J., Patterson, P., & Tuomaiainen, H. (2025). Universal mental health screening in schools: How acceptable is this to key stakeholders? A systematic review. <i>Journal of Child and Family Studies</i> , 34, 366–380. doi10.1007/s10826-025-03007-0	Provides the most current (2025), framework-grounded acceptability evidence for S1's acceptability/implementation-barriers structural gap, updating the 2019–2020 Anderson/Soneson evidence base.	VERIFIED verified 2026-08-25
Danielson, M. L., Kassab, H. D., Lee, M., Owens, J. S., Evans, S. W., Lipton, C., Charania, S., Young, H. E., Kubicek, L. F., Flory, K., & Cuffe, S. P. (2023). The utility of the Behavior Assessment System for Children-2 Behavioral and Emotional Screening System and Strengths and Difficulties Questionnaire in predicting mental disorders in the Project to Learn About Youth-Mental Health. <i>Psychology in the Schools</i> , 60(7), 2320–2341. doi10.1002/pits.22856	Directly substantiates S1's individual-actionability psychometric gaps for BESS/SDQ-style teacher-report screeners, especially their comparatively weak internalizing-disorder detection, a key structural gap the section highlights.	VERIFIED verified 2026-08-25

SECTION 11 · BIBLIOGRAPHY · CONTINUED

Section 1 · The Screener Landscape · 17 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Siceloff, E. R., Bradley, W. J., & Flory, K. (2017). Universal behavioral/emotional health screening in schools: Overview and feasibility. Report on Emotional & Behavioral Disorders in Youth, 17(2), 32–38.	Adds a resourcing and grade–band completion–rate lens to S1’s acceptability/feasibility structural gap, showing that even a successful pilot depended on support most districts do not have on their own.	VERIFIED verified 2026–08–25

Section 2 · From Self-Report To Authentic Voice · 13 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Bauer, M., & Zapata-Rivera, D. (2020). Cognitive foundations of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), Handbook of automated scoring: Theory into practice (pp. 13–28). Chapman and Hall/CRC.	Supports the Brief’s argument that construct–clarity and cognitive grounding, not model scale or black–box fit, are what make automated scoring trustworthy, the same argument Ober et al. make from the LLM–coding side in S2.	VERIFIED verified 2026–08–25
Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters, 140, 325–331. doi:10.1016/j.patrec.2020.11.008	Direct methodological citation for the CORAL ordinal–regression head named in Impacter Pathway’s own architecture (DeBERTa–v3–base + CORAL, per the live Model Card). Grounds the ‘discriminative ordinal pipeline, not generative sampling’ argument made in S2/S3, and the same head named in S7’s production–architecture discussion.	VERIFIED verified 2026–08–25
He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA–style pre–training with gradient–disentangled embedding sharing. arXiv:2111.09543.	The architecture–family paper for the DeBERTa–v3–base encoder Impacter Pathway’s scoring pipeline is built on, per the live, confirmed Model Card (impacterpathway.com/model–card). Supports the S3 model–facts framing.	VERIFIED verified 2026–08–25
IMPACTER Education Corp. (2026). IMPACTER Model Card, Version 6.0: CORAL Ordinal Regression Models.	The sole authorized source for any model–version accuracy figure appearing in the Brief.	INTERNAL verified 2026–08–25
Fisher, H., Jaffe, N. M., Pidvirny, K., Tierney, A. O., Vaidean, M. S., Dongre, P., & Webb, C. A. (2026). Language–based detection of depression with machine learning: Systematic review and meta–analysis. npj Digital Medicine, 9, Article 273. doi:10.1038/s41746–026–02448–1	Establishes the published accuracy ceiling for general–purpose language–based depression detection in the clinical NLP literature, useful context for framing a purpose–built, rubric–aligned scoring pipeline as a distinct, narrower measurement task rather than a clinical diagnostic claim.	VERIFIED verified 2026–08–25
Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116. doi:10.1002/liv.2.354	Frames the acoustic/paralinguistic branch of automated language–based mental health assessment, distinct from a text–based, rubric–scored approach, and shows how young this broader evidence base still is even for well–studied constructs like depression.	VERIFIED verified 2026–08–25
Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for depression using natural language processing: Literature review. Interactive Journal of Medical Research, 13, e55067. doi:10.2196/55067	Names the specific fairness, privacy, and interpretability risks that a multilingual, student–facing scoring system must design against.	VERIFIED verified 2026–08–25
Zapata–Rivera, D., & Forsyth, C. M. (2022). Learner modeling in conversation–based assessment. In Adaptive Instructional Systems: Adaptation Strategies and Methods (HCII 2022), Lecture Notes in Computer Science (Vol. 13332). Springer. doi:10.1007/978–3–031–05887–5_6	Establishes ETS’s conversation–based–assessment lineage as the closest existing large–testing–organization precedent for treating open–ended student language as a scorable measurement signal.	VERIFIED verified 2026–08–25
Lopez, A. A., Guzman–Orth, D., Zapata–Rivera, D., Forsyth, C. M., & Luce, C. (2021). Examining the accuracy of a conversation–based assessment in interpreting English learners’ written responses (Research Report No. RR–21–03). ETS Research Report Series. doi:10.1002/ets2.12315	Direct precedent for automated interpretation of open–ended, English–learner student language in an assessment context, supporting the case that language–based scoring is workable across linguistic backgrounds when the task is well–specified.	VERIFIED verified 2026–08–25
Ober, T. M., Zhang, S., Zapata–Rivera, D., Schroeder, N. L., & Botelho, A. F. (2026). Using LLMs to identify indicators of persistence from students’ dialogues with a pedagogical agent. Journal of Educational Data Mining, 18(1), 208–243. doi:10.5281/zenodo.18852441	Directly supports the argument that rubric–first, operationally precise construct definition is what a generative LLM–coding pipeline needs to be reliable, and by extension, why a discriminative, deterministic, frozen–and–versioned ordinal scoring architecture avoids the reproducibility problem this paper documents.	VERIFIED verified 2026–08–25
Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don’t always say what they think: Unfaithful explanations in chain–of–thought prompting. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).	Documents the class of instability and reproducibility risk inherent in generative, sampling–based LLM reasoning, the same risk category Ober et al. (2026) found in LLM–based construct coding, and the basis for distinguishing Impacter Pathway’s deterministic discriminative pipeline from generative LLM coding.	VERIFIED verified 2026–08–25
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.	Standard reference for transformer model–compression technique; relevant background for understanding tradeoffs between model size and scoring reliability discussed in the encoder–architecture literature this section engages.	VERIFIED verified 2026–08–25
He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding–enhanced BERT with disentangled attention. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).	Establishes the DeBERTa encoder lineage (disentangled attention over BERT/RoBERTa) that Impacter Pathway’s production DeBERTa–v3–base encoder descends from, distinguishing an encoder–plus–ordinal–head architecture from the generative decoder–only LLMs discussed elsewhere in this section.	VERIFIED verified 2026–08–25

Section 3 · Trustworthy Automated Scoring · 12 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring of constructed–response tasks. Educational Measurement: Issues and Practice, 31(1), 2–13. doi:10.1111/j.1745–3992.2011.00223.x	Supplies the evaluation vocabulary (agreement, degradation, fairness, consequences) that Impacter Pathway’s own promotion–gate framework (QWK ≥ 0.70, degradation ≤ 0.10, SMD < 0.10) is built on.	VERIFIED verified 2026–08–25

SECTION 11 · BIBLIOGRAPHY · CONTINUED

Section 3 · Trustworthy Automated Scoring · 12 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). Best practices for constructed-response scoring (Research Report No. RR-22-17). ETS Research Report Series. doi10.1002/ets2.12358	Corroborates Impacter Pathway's practice of governing human and automated scoring under one shared quality framework, and supports the Brief's evaluation-standards section with an ETS-sourced best-practices citation.	VERIFIED verified 2026-08-25
Wood, S., Yao, E., Haisfield, L., & Lottridge, S. (2021). Establishing standards of best practice in automated scoring (ACT Research Technical Brief). ACT, Inc.	A second independent testing-industry standards source (alongside Williamson/Xi/Breyer and AERA/APA/NCME) supporting the Brief's evaluation-standards framing; publicly hostable per the copyright rules.	VERIFIED verified 2026-08-25
Bauer, M., & Zapata-Rivera, D. (2020). Cognitive foundations of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), Handbook of automated scoring: Theory into practice (pp. 13-28). Chapman and Hall/CRC.	Supports the Brief's argument that construct-clarity and cognitive grounding, not model scale or black-box fit, are what make automated scoring trustworthy, the same argument Ober et al. make from the LLM-coding side in S2.	VERIFIED verified 2026-08-25
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.	The highest-authority umbrella standard the Brief can cite to frame Impacter Pathway's evaluation practices as consistent with the field's own professional testing standards.	VERIFIED verified 2026-08-25
Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. doi10.1177/001316446002000104	The foundational citation for the agreement-statistic vocabulary (kappa family, including QWK) used throughout Impacter Pathway's scoring-evaluation reporting.	VERIFIED verified 2026-08-25
Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research, 30(3), 411-433. doi10.1111/j.1468-2958.2004.tb00738.x	Second agreement-statistic foundation (alongside Cohen 1960) supporting Impacter Pathway's inter-rater and human-machine reliability reporting, especially useful where more than two raters or ordinal scales are involved.	VERIFIED verified 2026-08-25
Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters, 140, 325-331. doi10.1016/j.patrec.2020.11.008	Direct methodological citation for the CORAL ordinal-regression head named in Impacter Pathway's own architecture (DeBERTa-v3-base + CORAL, per the live Model Card). Grounds the 'discriminative ordinal pipeline, not generative sampling' argument made in S2/S3, and the same head named in S7's production-architecture discussion.	VERIFIED verified 2026-08-25
Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019) (pp. 1-10). Association for Computational Linguistics. doi10.18653/v1/W19-4401	Supplies the fuller fairness taxonomy against which Impacter Pathway's own promotion-gate [SMD] < 0.10 subgroup check can be located as one specific (overall-score-difference) fairness test among several, useful for scoping the Brief's fairness claims honestly rather than implying a single gate resolves the fairness question outright.	VERIFIED verified 2026-08-25
He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543.	The architecture-family paper for the DeBERTa-v3-base encoder Impacter Pathway's scoring pipeline is built on, per the live, confirmed Model Card (impacterpathway.com/model-card). Supports the S3 model-facts framing.	VERIFIED verified 2026-08-25
IMPACTER Education Corp. (2026). IMPACTER Model Card, Version 6.0: CORAL Ordinal Regression Models.	The sole authorized source for any model-version accuracy figure appearing in the Brief.	INTERNAL verified 2026-08-25
IMPACTER Education Corp. (2026). A stratified, human-in-the-loop sampling layer for continuous calibration of a production rubric-aligned ordinal scoring pipeline (Technical Note, Model Suite v6.0).	Supports S3's description of Impacter Pathway's continuous human-in-the-loop calibration practice as the operational complement to the academic evaluation-standards literature, and is S7's production-side evidence that human calibration is a standing weekly practice rather than a one-time validation event.	INTERNAL verified 2026-08-25

Section 4 · Process & Outcome Indicators · 8 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Scott, T. M., Gage, N. A., Hirn, R. G., Lingo, A. S., & Burt, J. (2019). An examination of the association between MTSS implementation fidelity measures and student outcomes. Preventing School Failure: Alternative Education for Children and Youth, 63(4), 308-316. doi10.1080/1045988X.2019.1605971	Direct empirical basis for BHPM's decision to track process/fidelity and outcome indicators as separate, independently reportable families rather than treating one as a proxy for the other.	VERIFIED verified 2026-08-25
Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for school-wide positive behavior support. Focus on Exceptional Children, 42(8), 1-14.	Supplies the framework-level (not single-practice) evidence-quality standard S4 uses to justify evaluating BHPM's own tiered indicator system as a whole rather than indicator-by-indicator.	VERIFIED verified 2026-08-25
McIntosh, K., Massar, M. M., Algozzine, R. F., George, H. P., Horner, R. H., Lewis, T. J., & Swain-Bradway, J. (2017). Technical adequacy of the SWPBIS Tiered Fidelity Inventory. Journal of Positive Behavior Interventions, 19(1), 3-13. doi10.1177/1098300716637193	Field-standard, psychometrically validated instrument that BHPM's process/fidelity indicators lean on wherever a site-level PBIS fidelity score is called for.	VERIFIED verified 2026-08-25
McIntosh, K., Herman, K., Bradshaw, C., & Simonsen, B. (2023). IES MTSS-B trial: Key takeaways for district and state leaders. Center on PBIS, University of Oregon.	The single independent RCT large enough to trust in this space, and the direct evidentiary basis for designing BHPM's outcome and contextual indicators to be disaggregated by subgroup and need tier from the start.	CANONICAL verified 2026-08-25
Substance Abuse and Mental Health Services Administration. (2026). Project AWARE: School campus mental health. U.S. Department of Health and Human Services.	Comparable federal reporting-category precedent for BHPM's service-utilization domain (partnerships, training reach, referral volume), borrowed as reporting categories only, never as evidence of causal program effect.	CANONICAL verified 2026-08-25
Cal. Welf. & Inst. Code § 5961.4 (2024).	The statutory basis for the CYBHI Fee Schedule Program, and the legal specificity behind BHPM's service-utilization indicator tying screening activity to a reimbursable service touchpoint.	CANONICAL verified 2026-08-25
California Department of Health Care Services. (2026). Children and Youth Behavioral Health Initiative (CYBHI) Fee Schedule Program.	The in-state, statewide multi-payer reporting infrastructure BHPM's own service-utilization indicators can eventually connect to as a district's CYBHI cohort status and billing infrastructure come online.	CANONICAL verified 2026-08-25

SECTION 11 · BIBLIOGRAPHY · CONTINUED

Section 4 · Process & Outcome Indicators · 8 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
American Medical Association. (2026). CPT code 96127: Brief emotional/behavioral assessment, with scoring and documentation, per standardized instrument. Current Procedural Terminology.	The specific billing-code detail that ties a BHPM screening event to a documented, reimbursable service-utilization touchpoint wherever a site's billing infrastructure already supports it.	CANONICAL verified 2026–08–25

Section 5 · Implementation Science · 6 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). Implementation research: A synthesis of the literature (FMHI Publication #231). National Implementation Research Network, University of South Florida, Louis de la Parte Florida Mental Health Institute.	Supplies the stage/component framework the Brief uses to argue that BHPM's readiness-to-deployment-to-TTA sequence (D1 to D2 to D3) is a deliberate, evidence-grounded implementation pathway rather than an ad hoc rollout.	CANONICAL verified 2026–08–25
Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. <i>Administration and Policy in Mental Health and Mental Health Services Research</i> , 38(2), 65–76. doi10.1007/s10488-010-0319-7	Gives the Metrics Brief a validated eight-domain taxonomy for organizing process-fidelity indicators separately from outcome indicators within the composite index and evidence tables.	VERIFIED verified 2026–08–25
Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). <i>Learning to improve: How America's schools can get better at getting better</i> . Harvard Education Press.	Provides the direct literature basis for BHPM's D4/D5 Collaborative Learning Group design, framing the quarterly learning-group cycle as PDSA-style networked improvement rather than a one-off convening.	CANONICAL verified 2026–08–25
Baffsky, R., Ivers, R., Cullen, P., Wang, J., McGillivray, L., & Torok, M. (2023). Strategies for enhancing the implementation of universal mental health prevention programs in schools: A systematic review. <i>Prevention Science</i> , 24(2), 337–352. doi10.1007/s11121-022-01434-9	Directly supports the case that BHPM's continuous platform-analytics approach to monitoring and feedback, rather than annual recall surveys, sits on the strongest evidence stream this review identifies for school-based behavioral-health implementation.	VERIFIED verified 2026–08–25
Deming, W. E. (1986). <i>Out of the crisis</i> . Massachusetts Institute of Technology, Center for Advanced Engineering Study.	Establishes the improvement-science lineage that BHPM's continuous, cycle-based process-evaluation design draws on, rather than treating PDSA as if it originated with the education-sector literature alone.	CANONICAL verified 2026–08–25
Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. <i>Implementation Science</i> , 4, Article 50. doi10.1186/1748-5908-4-50	Gives BHPM's Deliverable 1 site readiness assessment a validated, construct-level checklist for the organizational-context factors CFIR's inner-setting domain identifies as separable from technical capacity, complementing rather than replacing the Fixsen driver framework already anchoring this section.	VERIFIED verified 2026–08–25

Section 6 · Composite Index Construction · 5 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
OECD, & European Commission Joint Research Center. (2008). <i>Handbook on constructing composite indicators: Methodology and user guide</i> . OECD Publishing. doi10.1787/9789264043466-en	The direct, step-by-step methodological framework the Behavioral Health Performance Index's entire construction (sub-indices, normalization, weighting, aggregation, banding, versioning) is built on.	CANONICAL verified 2026–08–25
McDonald, R. P. (1999). <i>Test theory: A unified treatment</i> . Lawrence Erlbaum Associates. doi10.4324/9781410601087	The psychometric basis for BHPI's commitment to report composite and sub-index reliability using omega rather than Cronbach's alpha, given that each domain sub-index mixes heterogeneous indicator types.	VERIFIED verified 2026–08–25
Seastrom, M. (2010). Statistical methods for protecting personally identifiable information in aggregate reporting (NCES 2011–603; SLDS Technical Brief 3). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.	The federal basis for BHPI's default minimum reporting-group size of ten students plus complementary suppression, set with SCCOE data governance.	VERIFIED verified 2026–08–25
California Department of Education. (2025). <i>California School Dashboard technical guide: 2025 Dashboard</i> .	The in-state design precedent BHPI's own performance-banding approach adapts, so LEA teams read behavioral health performance in the same visual grammar they already use for academic and climate indicators.	CANONICAL verified 2026–08–25
Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. <i>Journal of the Royal Statistical Society: Series A (Statistics in Society)</i> , 168(2), 307–323. doi10.1111/j.1467-985X.2005.00350.x	Supplies the specific statistical technique (variance-based/Sobol-type decomposition) behind the general 'mandatory sensitivity analysis' requirement the OECD/JRC Handbook states, and grounds the Index Methodology page's plan to report how much of BHPI's score and rank uncertainty is attributable to the weight set specifically versus the normalization or aggregation choice, rather than reporting one undifferentiated robustness number.	VERIFIED verified 2026–08–25

Section 7 · Research Lineage · 12 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. <i>Pattern Recognition Letters</i> , 140, 325–331. doi10.1016/j.patrec.2020.11.008	Direct methodological citation for the CORAL ordinal-regression head named in Impacter Pathway's own architecture (DeBERTa-v3-base + CORAL, per the live Model Card). Grounds the 'discriminative ordinal pipeline, not generative sampling' argument made in S2/S3, and the same head named in S7's production-architecture discussion.	VERIFIED verified 2026–08–25
IMPACTER Education Corp. (2026). <i>IMPACTER Model Card, Version 6.0: CORAL Ordinal Regression Models</i> .	The sole authorized source for any model-version accuracy figure appearing in the Brief.	INTERNAL verified 2026–08–25

SECTION 11 · BIBLIOGRAPHY · CONTINUED

Section 7 · Research Lineage · 12 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
IMPACTER Education Corp. (2026). A stratified, human-in-the-loop sampling layer for continuous calibration of a production rubric-aligned ordinal scoring pipeline (Technical Note, Model Suite v6.0).	Supports S3's description of Impacter Pathway's continuous human-in-the-loop calibration practice as the operational complement to the academic evaluation-standards literature, and is S7's production-side evidence that human calibration is a standing weekly practice rather than a one-time validation event.	INTERNAL verified 2026–08–25
Ober, T. M., Zhang, S., Zapata-Rivera, D., Schroeder, N. L., & Botelho, A. F. (2026). Using LLMs to identify indicators of persistence from students' dialogues with a pedagogical agent. <i>Journal of Educational Data Mining</i> , 18(1), 208–243. doi10.5281/zenodo.18852441	Directly supports the argument that rubric-first, operationally precise construct definition is what a generative LLM-coding pipeline needs to be reliable, and by extension, why a discriminative, deterministic, frozen-and-versioned ordinal scoring architecture avoids the reproducibility problem this paper documents.	VERIFIED verified 2026–08–25
Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In <i>Advances in Neural Information Processing Systems 36 (NeurIPS 2023)</i> .	Documents the class of instability and reproducibility risk inherent in generative, sampling-based LLM reasoning, the same risk category Ober et al. (2026) found in LLM-based construct coding, and the basis for distinguishing Impacter Pathway's deterministic discriminative pipeline from generative LLM coding.	VERIFIED verified 2026–08–25
Arnold, A., Liu, Y., & Abe, N. (2007). Temporal causal modeling with graphical Granger methods. In <i>Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining</i> (pp. 66–75). Association for Computing Machinery. doi10.1145/1281192.1281203	Documents the specific causal-inference toolkit behind Impacter Pathway's Senior Scientific Advisor and scoring-pipeline architect, framed as the planned analytic foundation for BHPM's Year 2 upstream-lever and dosage analysis with Stanford Medicine, not as evidence about the screener itself.	VERIFIED verified 2026–08–25
Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. <i>Econometrica</i> , 37(3), 424–438. doi10.2307/1912791	The direct intellectual ancestor of the graphical Granger-causality methods in Arnold, Liu & Abe (2007): documents that Impacter Pathway's Senior Scientific Advisor's causal-inference toolkit traces to a specific, checkable methodological origin rather than being a generic 'machine learning' credential.	VERIFIED verified 2026–08–25
IMPACTER Education Corp. (2026). From self-report to student voice: A guide to performance-based behavioral health screening (Version 2.1).	The plainer-language, partner-facing statement of the same design choice S7 documents academically: rubric-aligned scoring against authentic student language rather than a fixed self-report checklist.	INTERNAL verified 2026–08–25
IMPACTER Education Corp. (2026). A human-friendly guide to IMPACTER's machine learning model: A resource for our partners.	The organization's own plain-language statement of the human-leads/machine-learning-scales principle that S7 closes its research-lineage argument with.	INTERNAL verified 2026–08–25
Hirshberg, M. J., Colaianne, B. A., Greenberg, M. T., Inkelas, K. K., Davidson, R. J., Germano, D., Dunne, J. D., & Roeser, R. W. (2022). Can the academic and experiential study of flourishing improve flourishing in college students? A multi-university study. <i>Mindfulness</i> , 13(9), 2243–2256. doi10.1007/s12671-022-01952-1	Design rationale, not validity evidence, for why Impacter Pathway's prompts ask students to reflect in their own words rather than select from a fixed checklist.	VERIFIED verified 2026–08–25
Pennebaker, J. W., & Beall, S. K. (1986). Confronting a traumatic event: Toward an understanding of inhibition and disease. <i>Journal of Abnormal Psychology</i> , 95(3), 274–281. doi10.1037/0021-843X.95.3.274	The earlier link in the same reflective-practice research lineage Hirshberg et al. (2022) extends: evidence that structured reflection on one's own experience is itself a mechanism with measurable effects, not merely a delivery format: design rationale for Impacter Pathway's open-ended prompts, not validity evidence for the screener.	VERIFIED verified 2026–08–25
Saab, B. J., Georgiou, J., Nath, A., Lee, F. J. S., Wang, M., Michalon, A., Liu, F., Mansuy, I. M., & Roder, J. C. (2009). NCS-1 in the dentate gyrus promotes exploration, synaptic plasticity, and rapid acquisition of spatial memory. <i>Neuron</i> , 63(5), 643–656. doi10.1016/j.neuron.2009.08.014	Cited strictly as Impacter Pathway's stated founding neurobiological inspiration for taking the biological basis of exploratory behavior seriously as a construct worth measuring, explicitly NOT validity evidence for the screener and not evidence about human behavioral health outcomes; this boundary must be preserved in any downstream reuse.	VERIFIED verified 2026–08–25

Cited Outside The Seven Sections · 2 Sources

SOURCE	WHAT IT ESTABLISHES	PROVENANCE
Elo, A. E. (1978). <i>The rating of chessplayers, past and present</i> . Arco Publishing.	Flagged in the Index Methodology page as a literature worth considering for the BHPI change-axis calculation specifically (weighting year-over-year movement by measurement confidence rather than by simple subtraction), not as part of the core four-domain composite construction itself.	CANONICAL verified 2026–08–25
Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. <i>Journal of the Royal Statistical Society: Series C (Applied Statistics)</i> , 48(3), 377–394. doi10.1111/1467-9876.00159	Cited alongside Elo (1978) in the Index Methodology page as the more rigorous, uncertainty-aware paired-comparison alternative worth considering for BHPI's year-over-year change-axis calculation.	VERIFIED verified 2026–08–25

How to read provenance

VERIFIED means the entry was resolved to a live DOI, publisher record or open-access record on the date shown. INTERNAL means an Impacter Pathway work product. CANONICAL means the approved Model Card: the single source for any model-performance figure quoted anywhere in this record.

SECTION 12 · ARTIFACTS

The Core Documents Behind The Brief

The core documents the brief rests on: contract and workplan records, scoring and model documentation, screening guides, the independent peer-reviewed research, and the Deliverable 1 readiness artifacts.

Contract & Workplan

DOCUMENT	WHAT IT IS	PROVENANCE
Implementation Workplan Crosswalk published record	The master crosswalk mapping every BHPM contract deliverable and activity to its owner, timeline, and Commission reporting domain across the nine-deliverable Implementation Workplan. Published on this site as a readable record.	INTERNAL

Scoring & Model Documentation

DOCUMENT	WHAT IT IS	PROVENANCE
Technical Note v6.x · Stratified Human-in-the-Loop Calibration	Impacter Pathway's internal technical note on the stratified, weekly human-in-the-loop sampling layer that calibrates the scoring pipeline: sampling design across attribute, score band, prompt, and item type, plus drift monitoring and retraining governance.	INTERNAL
Human-Friendly Guide to IMPACTER's ML Model	A plain-language walkthrough of how Impacter Pathway's machine learning model reads a student's open-ended response and produces a score, written for non-technical readers.	INTERNAL
IMPACTER Model Card v6.0	The canonical, published model card for Impacter Pathway's Model Suite v6.0: a fine-tuned DeBERTa-v3-base encoder with a CORAL ordinal regression head (184M parameters), averaging 0.918 QWK, 77.3% exact-match accuracy, and 98.0% adjacent (+/-1) accuracy against human raters across the eight durable-skill attributes, with promotion gates of QWK ≥ 0.70 , degradation ≤ 0.10 , and SMD ≤ 0.10 .	CANONICAL

Screening & Assessment Guides

DOCUMENT	WHAT IT IS	PROVENANCE
Behavioral Health Screening Guide (BHPM)	A guide to performance-based behavioral health screening through authentic student voice, developed for the BHPM project and its participating LEAs.	INTERNAL

Independent Peer-Reviewed Research

DOCUMENT	WHAT IT IS	PROVENANCE
Arnold, Liu & Abe (2007) · Temporal Causal Modeling With Graphical Granger Methods ↗	Arnold, A., Liu, Y., & Abe, N. (2007). Temporal causal modeling with graphical Granger methods. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 66–75. https://doi.org/10.1145/1281192.1281203 . Introduces Lasso-Granger for learning temporal dependency among multiple time series via variable selection.	VERIFIED
Ober et al. (2026) · Using LLMs To Identify Indicators Of Persistence	Ober, T., 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. Open access; supplemental materials at the OSF project doi.org/10.17605/OSF.IO/S85CK .	VERIFIED
Hirshberg, Colaianne et al. (2022) · Flourishing In College Students ↗	Hirshberg, M. J., Colaianne, B. A., Greenberg, M. T., Inkelas, K. K., Davidson, R. J., Germano, D., Dunne, J. D., & Roeser, R. W. (2022). Can the academic and experiential study of flourishing improve flourishing in college students? A multi-university study. Mindfulness, 13(9), 2243–2256. https://doi.org/10.1007/s12671-022-01952-1 . Open access via PMC.	VERIFIED
Saab et al. (2009) · NCS-1 In The Dentate Gyrus, Neuron ↗	Saab, B. J., Georgiou, J., Nath, A., Lee, F. J., Wang, M., Michalon, A., Liu, F., Mansuy, I. M., & Roder, J. C. (2009). NCS-1 in the dentate gyrus promotes exploration, synaptic plasticity, and rapid acquisition of spatial memory. Neuron, 63(5), 643–656. https://doi.org/10.1016/j.neuron.2009.08.014 . Paywalled (Elsevier); linked to the publisher DOI record only, not rehoused.	VERIFIED

D1 Readiness Artifacts

DOCUMENT	WHAT IT IS	PROVENANCE
D1 · Site Readiness Assessment (Google Doc) ↗	The live Site Readiness Assessment tracking readiness criteria and current status by site across the six participating LEAs, Impacter Pathway's contribution to Deliverable 1.	INTERNAL
D1 · Cohort Readiness Tracker (Google Sheet) ↗	The live Cohort Readiness Tracker recording enrollment and readiness status by school across the eighteen participating schools, Impacter Pathway's contribution to Deliverable 1.	INTERNAL

COLOPHON

How To Cite This Brief

Santa Clara County Office of Education & Impacter Education Corp. (2026). *Deliverable 2: Indicators, Outcomes & Metrics Brief*. Behavioral Health Performance Management, BHSOAC Contract #25BHSOAC019.
bhpm.care/deliverables/d2

Where The Figures Come From

- The record itself: every summary on these sheets is a section's own closing callout or opening paragraph
- The index tables: the five weights, the four bands, the renormalization rule and the six-competency rubric
- The indicator table: thirteen candidate indicators with family, domain and feasibility
- The screener comparison table: eight instruments by informant, format and actionability
- The source registry: sixty-six entries with provenance and verification date
- The earlier working design: the diagram data for the evaluation criteria, the staged sequence, the lineage and the crosswalk pattern

The Record Behind This Brief

The twelve sections are published at bhpm.care with the sources beside the prose. The [Deliverable 2 reader](#) carries the full text; the [earlier working design](#) keeps the interactive evidence tables and diagrams; the [resource library](#) holds the documents.

12	66
Sections Published	Sources In The Registry

What Is Published Underneath This Brief

RECORD	WHAT IT HOLDS	ADDRESS
Deliverable 2 Reader	The twelve sections of this brief, each as its own page, with the sources cited beside the prose	bhpm.care/deliverables/d2
The Earlier Working Design	The seven literature sections with their interactive evidence tables, diagrams and source cards	bhpm.care/archive/d2-v1
Implementation Workplan Crosswalk	Every contract workplan item against the deliverable that answers it	bhpm.care/record/crosswalk
Deliverable 1 Reader	The Recruitment Outcomes Summary: the cohort this brief is written for	bhpm.care/deliverables/d1
Resource Library	The contract, the agreements, the screening guides and this brief as a PDF	bhpm.care/resources

LED BY THE SANTA CLARA COUNTY OFFICE OF EDUCATION

PRIME

DELIVERY

TA

RESEARCH

POLICY

EVALUATION