Composite Health Score: A Statistical Framework for Measuring and Tracking Organizational Data Quality in Jira Environments

Authors: [Organization Name] **Version:** 1.1 **Date:** January 2026

Abstract

Norm-referenced performance measurement systems face a fundamental challenge: when all units improve uniformly, percentile rankings remain unchanged, creating the illusion of stagnation despite genuine progress. This paper presents the Composite Health Score (CHS), a statistical framework designed to measure team health in software development environments while addressing this invariance problem.

The CHS methodology combines three complementary components: the Current State Score (CSS), which measures absolute position against fixed baseline norms; the Trajectory Score (TRS), which captures improvement momentum using effect-size methodology; and the Peer Growth Score (PGS), which contextualizes growth relative to teams with similar starting positions. Drawing on established approaches from educational measurement—particularly Student Growth Percentiles (Betebenner, 2009)—the framework adapts these methods for organizational assessment contexts while remaining computationally feasible in constrained serverless environments.

Key methodological features include fixed baseline standardization to enable longitudinal comparability, Cohen's d effect sizes for trajectory measurement, and Empirical Bayes shrinkage to stabilize estimates from small peer groups. The framework provides analytic standard errors and 90% confidence intervals for uncertainty quantification, with explicit documentation of approximations and their limitations.

The CHS addresses the needs of organizations investing in process improvement initiatives by making genuine improvement visible in scores, even when all teams improve together. Complementary to the Composite Progress Score (CPS) for measuring change between assessment points, the CHS provides the "before" and "after" measurements that enable meaningful progress tracking. The methodology has been designed for production deployment in enterprise software tools, balancing statistical rigor with practical implementation constraints.

---

1. Introduction

1.1 Context and Motivation

Software development teams increasingly rely on work management tools such as Jira to coordinate complex projects, track progress, and manage dependencies. As organizations scale, questions arise about team effectiveness: Are teams following good practices? How do practices compare across teams? Are improvement initiatives working?

Unlike standardized testing where external benchmarks exist, there is no universal standard for "good" Jira hygiene. Optimal practices for estimation coverage, comment frequency, or work-in-progress limits vary by organizational context, team size, project type, and development methodology. A small startup team working on a greenfield product requires different practices than a large enterprise team maintaining critical infrastructure.

In the absence of absolute standards, **norm-referenced assessment** provides a principled alternative. Rather than comparing teams to an external benchmark, norm-referenced approaches compare teams to each other within the same organizational context. This approach has extensive precedent in educational testing, employee evaluation, and performance management (Glaser, 1963; Popham, 1978).

Norm-referenced assessment works well for initial positioning. A team learning that they rank at the 25th percentile for estimation coverage receives actionable information: they are below average relative to peers and may benefit from improvement. The challenge emerges when measuring progress over time.

1.2 The Progress Measurement Problem

Percentile rankings depend only on ordinal position. If all teams improve by the same amount, all ordinal relationships are preserved, and all percentiles remain unchanged—despite genuine improvement across the organization.

Formal Statement: Let $X_t^{(0)}$ and $X_t^{(1)}$ denote team t$'s scores at times 0 and 1. If $X_t^{(1)} = X_t^{(0)} + c for all teams (uniform additive shift), then $P_t^{(1)} = P_t^{(0)}$ for all teams, where $P_t^{(k)}$ denotes the percentile rank at time k.

This invariance property creates practical problems. Organizations that invest in improvement initiatives may see genuine metric improvements across all teams, yet percentile-based dashboards show "no change." Stakeholders reviewing dashboards see static numbers and question whether the initiative had any effect. Teams that worked hard to improve feel their efforts are invisible. The measurement system, rather than supporting improvement, actively undermines it.

Consider a concrete example: An organization implements estimation training across all 50 teams. After three months, estimation coverage increases from an average of 65% to 80% across all teams. Under pure percentile scoring, every team's percentile remains exactly the same. The training appears to have had no effect, when in reality all teams improved substantially.

1.3 Scope and Objectives

This paper specifies a methodology to:

1. **Detect genuine improvement** even when all teams improve together, by incorporating absolute measurement against fixed baseline norms

2. **Contextualize improvement** relative to peers with similar starting points, so that a team improving from the 30th to 50th percentile receives appropriate credit given their baseline
3. **Quantify uncertainty** in health estimates through analytic standard errors and confidence intervals
4. **Remain computationally feasible** in constrained serverless runtime environments where iterative optimization and resampling methods are impractical

The framework does not claim to enable causal inference—improvement may result from factors other than deliberate intervention—nor does it claim optimal precision for individual scores. Standard errors are provided, and small score differences should not be over-interpreted.

1.4 Contributions and Paper Organization

This paper makes the following contributions:

1. **Multi-component health score** addressing the percentile invariance problem through CSS, TRS, and PGS components
2. **Fixed baseline norm approach** for CSS enabling longitudinal comparability without continuous recalibration
3. **Effect-size trajectory measurement** for TRS capturing absolute magnitude of change
4. **Empirical Bayes shrinkage** for PGS stabilizing estimates from small peer groups
5. **Complete computational specification** suitable for implementation in constrained serverless environments
6. **Complementary CPS framework** for measuring progress between assessment points

The paper is organized as follows: Section 2 reviews relevant literature on progress measurement, growth models, and shrinkage estimation. Section 3 presents the complete methodology including CSS, TRS, PGS, and composite CHS calculation. Section 4 justifies the interpretation thresholds and sensitivity analysis approach. Section 5 provides comprehensive scenario analysis with worked examples. Section 6 discusses implementation considerations. Section 7 addresses limitations and future work. Section 8 concludes.

---

2. Literature Review

2.1 Norm-Referenced Assessment

Glaser (1963) established the fundamental distinction between criterion-referenced and norm-referenced measurement, demonstrating that the latter is appropriate when “the standard is the performance of some reference group” rather than an external benchmark. In software development practices, there is no universal standard for what constitutes “good” estimation coverage or comment frequency—optimal values vary by team size, project type, and organizational culture.

Popham (1978) extended this framework, noting that norm-referenced approaches excel at differentiating performers within a population but struggle to capture uniform

improvement. This limitation motivates the CHS multi-component design: pure percentile ranking (norm-referenced) is combined with fixed-baseline standardization (quasi-criterion-referenced via CSS) and conditional peer comparison (PGS).

2.2 Simple Gain Scores and Their Limitations

The naive approach to measuring change is the simple gain score:

$$G_t = S_t^{(1)} - S_t^{(0)}$$

Lord (1956) identified fundamental problems with this approach, noting that "a difference score between two fallible measures is likely to be more fallible than either... gain scores tend to be negatively correlated with initial status." Teams starting at extreme positions—very low or very high scores—will naturally regress toward the mean on subsequent measurement, appearing to "improve" or "decline" independently of any genuine change.

Cronbach and Furby (1970) influentially argued that "investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways." Their concerns were not merely theoretical: regression artifacts confound interpretation in any longitudinal study using simple differences.

Campbell and Kenny (1999) provided comprehensive treatment of these regression artifacts, demonstrating their ubiquity in longitudinal research. However, Rogosa and Willett (1985) offered a more optimistic view, showing that growth can be reliably measured when based on multiple observations and appropriate modeling.

2.3 Student Growth Percentiles

Student Growth Percentiles (SGP), introduced by Betebenner (2009), provide the theoretical foundation for conditional progress measurement. The key insight reframes the progress question: rather than asking "what percentile is this team at now?", SGP asks "among teams that started at a similar level, what percentile of growth did this team achieve?"

Formal Definition:

$$SGP_t = P(S^{(1)} \leq s_t^{(1)} \mid S^{(0)} = s_t^{(0)}) \times 100$$

This conditioning on prior achievement addresses regression to the mean and provides fair comparison across different starting positions. A team improving from the 30th to 50th percentile achieves more impressive growth than a team moving from the 80th to 85th, and SGP captures this distinction.

The original SGP methodology employs quantile regression (Koenker & Bassett, 1978) to model the conditional distribution of outcomes. Castellano and Ho (2013) provided practitioner guidance specifically recommending discrete baseline grouping for sample sizes below 100—the approach CHS adopts—as both statistically appropriate and computationally trivial.

2.4 Effect Size Measures

Effect sizes provide magnitude-invariant measurement that addresses the percentile invariance problem. Cohen (1988) established the d statistic as the standard for expressing differences in standard deviation units:

$$d = \frac{X^{(1)} - X^{(0)}}{\sigma_{baseline}}$$

with conventions for interpretation: $|d| = 0.2$ (small), $|d| = 0.5$ (medium), $|d| = 0.8$ (large).

Unlike percentiles, which depend only on ordinal position, effect sizes capture actual magnitude of change. If all teams improve by 0.5 standard deviations, all effect sizes reflect this genuine improvement rather than showing "no change" as percentiles would.

Cumming (2014) advocated for effect sizes as the core of "the new statistics," emphasizing their greater interpretability compared to p-values. Glass, McGaw, and Smith (1981) demonstrated that effect sizes can be meaningfully aggregated, providing theoretical support for combining indicator-level effect sizes into composite measures.

2.5 Empirical Bayes Methods

Empirical Bayes shrinkage addresses the instability of estimates from small groups. Morris (1983) provided the theoretical foundation for parametric empirical Bayes inference, demonstrating how information can be pooled across groups to improve estimation in each group individually.

Efron and Morris (1975) showed that shrinkage estimators dominate separate estimation even when the groups being combined have no substantive relationship—the Stein paradox. The mechanism is shrinkage: extreme estimates from small samples are "pulled" toward the overall mean, with the degree of shrinkage inversely related to sample size.

For the PGS component, when a baseline group contains only 5-10 teams, raw percentile estimates are highly variable. Shrinkage toward 50 (the grand mean percentile) stabilizes these estimates, accepting some bias in exchange for substantial variance reduction. Raudenbush and Bryk (2002) extended these ideas to hierarchical data structures, providing the variance-component estimation methods that inform parameter calibration.

---

3. Methodology

3.1 Overview and Notation

The Composite Health Score combines three components:

Component	Weight	Purpose

CSS (Current State Score)	50%	Where the team is NOW relative to fixed baseline norms
TRS (Trajectory Score)	35%	How the team is TRENDING within the assessment period
PGS (Peer Growth Score)	15%	Growth relative to teams with similar starting positions

Core Notation:

Symbol	Definition
$T = \{t_1, \ldots, t_n\}$	Set of n teams
$D = \{d_1, \ldots, d_D\}$	Set of D dimensions (default: 14)
$I_d = \{i_{d,1}, \ldots, i_{d,m_d}\}$	Set of m_d indicators in dimension d
$I = \bigcup_{d=1}^{D} I_d$	Complete set of all indicators
$X_{t,d,i}^{(k)}$	Raw value of indicator i in dimension d for team t at time k
$\delta_{d,i} \in \{+1, -1\}$	Directionality of indicator i in dimension d
$\tilde{X}_{t,d,i}^{(k)}$	Direction-adjusted value: $\tilde{X}_{t,d,i}^{(k)} = \delta_{d,i} \cdot X_{t,d,i}^{(k)}$
w_d	Weight for dimension d, where $\sum_{d=1}^{D} w_d = 1$
$w_{d,i}$	Weight for indicator i within dimension d, where $\sum_{i=1}^{m_d} w_{d,i} = 1$
$\mu_{\text{baseline},d,i}$, $\sigma_{\text{baseline},d,i}$	Fixed baseline norms for indicator (d, i)
$\bar{\rho}$	Average pairwise correlation among indicators

Default Hierarchical Equal Weighting:

- Dimension weights: $w_d = \frac{1}{D}$ (equal across all dimensions)
- Indicator weights within dimension: $w_{d,i} = \frac{1}{m_d}$ (equal within each dimension)

3.2 Current State Score (CSS)

Purpose: Measure where the team is RIGHT NOW relative to fixed baseline population norms.

Key Design Decision: CSS uses **fixed baseline norms** established during initial calibration, not the current peer distribution. This ensures scores are comparable over time and not affected by shifting peer distributions.

3.2.1 Calculation Steps

Step 1: Direction Adjustment

For each indicator i, apply directionality so that higher values always represent better performance:

$$\tilde{X}_{t,i} = d_i \cdot X_{t,i}$$

where $d_i = +1$ if higher is better, $d_i = -1$ if lower is better.

Step 2: Winsorization

Before computing z-scores, winsorize at the 2nd and 98th percentiles to limit the influence of extreme values:

$$\tilde{X}_{t,i} \leftarrow \text{clip}\left(\tilde{X}_{t,i}, P_2(\tilde{X}_i), P_{98}(\tilde{X}_i)\right)$$

Step 3: Standardization

For each indicator i:

$$z_i = \frac{\tilde{X}_{t,i} - \mu_{\text{baseline},i}}{\sigma_{\text{baseline},i}}$$

Step 4: Hierarchical Aggregation with Variance Adjustment

With hierarchical weighting, aggregation proceeds in two stages:

Stage 1: Dimension-level aggregation

For each dimension d, compute the dimension score as the weighted sum of indicator z-scores:

$$z_d = \sum_{i=1}^{m_d} w_{d,i} \cdot z_{d,i}$$

Stage 2: Overall aggregation

The raw CSS is the weighted sum of dimension scores:

$$\text{CSS}_{\text{raw}} = \sum_{d=1}^{D} w_d \cdot z_d$$

With equal dimension weights ($w_d = \frac{1}{D}$) and equal indicator weights within dimensions ($w_{d,i} = \frac{1}{m_d}$):

$$\text{CSS}_{\text{raw}} = \frac{1}{D} \sum_{d=1}^{D} \frac{1}{m_d} \sum_{i=1}^{m_d} z_{d,i}$$

Variance calculation:

The variance of the aggregate, accounting for indicator correlations, is:

$$\text{Var}(\text{CSS}_{\text{raw}}) = \sum_{d=1}^{D} w_d^2 \cdot \text{Var}(z_d) + 2 \sum_{d<d'} w_d w_{d'} \cdot \text{Cov}(z_d, z_{d'})$$

With the simplifying assumption of common within-dimension correlation $\bar{\rho}$ and independence across dimensions:

$$\text{Var}(\text{CSS}_{\text{raw}}) \approx \sum_{d=1}^{D} w_d^2 \cdot \left[\sum_{i=1}^{m_d} w_{d,i}^2 \cdot (1 - \bar{\rho}) + \bar{\rho} \right]$$

This formula applies when weights sum to 1.0 at each level and indicators have unit variance after standardization.

Step 5: Scaling

To achieve a target standard deviation of 15 points (consistent with T-score conventions):

$$k_{\text{CSS}} = \frac{15}{\sqrt{\text{Var}(\text{CSS}_{\text{raw}})}}$$

$$\text{CSS} = 50 + k_{\text{CSS}} \cdot \text{CSS}_{\text{raw}}$$

Step 6: Bounding

Bound to $[5, 95]$ to avoid extreme scores:

$$\text{CSS} = \text{clip}(\text{CSS}, 5, 95)$$

3.2.2 Standard Error of CSS

The standard error, accounting for sampling variability and indicator correlation:

$$\text{SE}(\text{CSS}_{\text{raw}}) = \sqrt{\sum_{i=1}^{m} w_i^2 \cdot \frac{2}{n-1}} \cdot \sqrt{1 + \bar{\rho}(m-1)}$$

$$\text{SE}(\text{CSS}) = k_{\text{CSS}} \cdot \text{SE}(\text{CSS}_{\text{raw}})$$

where n is the number of observations (sprints/weeks) used to calculate each indicator.

3.2.3 Interpretation

- CSS = 50: Average current state relative to baseline norms
- CSS = 65: Current state is 1 SD above baseline average
- CSS = 35: Current state is 1 SD below baseline average

3.3 Trajectory Score (TRS)

Purpose: Measure how the team is TRENDING within the assessment period using effect-size methodology.

3.3.1 Calculation Steps

Step 1: Period Segmentation

Divide the assessment period (typically 6 months) into early and recent periods:

- **Early period:** First 4-6 time periods (sprints/weeks)
- **Recent period:** Last 4-6 time periods

Step 2: Effect Size per Indicator

For each indicator i, compute Cohen's d:

$$\text{trajectory}_i = \frac{\bar{X}_{\text{recent},i} - \bar{X}_{\text{early},i}}{\sigma_{\text{pooled},i}}$$

where σ_{pooled} is the pooled standard deviation across all periods.

Step 3: Winsorization

At indicator level, clip extreme effect sizes:

$$\text{trajectory}_i^* = \text{clip}(\text{trajectory}_i, -3, +3)$$

Step 4: Aggregation

$$\text{TRS}_{\text{raw}} = \sum_{i=1}^{m} w_i \cdot \text{trajectory}_i^*$$

Step 5: Aggregate-Level Winsorization

$$\text{TRS}_{\text{raw}}^* = \text{clip}(\text{TRS}_{\text{raw}}, -4.5, +4.5)$$

Step 6: Scaling

$$\text{TRS} = 50 + 10 \cdot \text{TRS}_{\text{raw}}^*$$

$$\text{TRS} = \text{clip}(\text{TRS}, 5, 95)$$

3.3.2 Standard Error of TRS

$$\text{SE}(\text{TRS}_{\text{raw}}) = \sqrt{\sum_{i=1}^{m} w_i^2 \cdot \frac{2}{n_{\text{periods}} - 1}} \cdot \sqrt{1 + \bar{\rho}(m-1)}$$

$$\text{SE}(\text{TRS}) = 10 \cdot \text{SE}(\text{TRS}_{\text{raw}})$$

3.3.3 Interpretation

- TRS = 50: Stable, no significant change during assessment period
- TRS = 70: Strong positive trajectory (+2 SD improvement rate)
- TRS = 30: Concerning negative trajectory (-2 SD decline rate)

3.4 Peer Growth Score (PGS)

Purpose: Compare trajectory to peers who started at a similar level, following the Student Growth Percentile paradigm.

3.4.1 Baseline Group Formation

Based on CSS at the START of the assessment period:

Sample Size (n)	Grouping
$n \geq 50$	Deciles (10 groups)
$30 \leq n < 50$	Quintiles (5 groups)
$20 \leq n < 30$	Quartiles (4 groups)
$n < 20$	Do not compute PGS

3.4.2 Minimum Group Size and Merge Procedure

Requirement: Each baseline group must contain at least 5 teams.

Merge Procedure:

```
WHILE any group has < 5 teams:
  smallest = group with minimum team count

  IF smallest is the lowest group: merge with next higher
  ELSE IF smallest is the highest group: merge with next lower
  ELSE: merge toward distribution center

  Log all merges for audit
```

3.4.3 Raw Percentile Calculation

Within each baseline group g:

$$\text{PGS}_{\text{raw}} = \frac{\text{rank}(\text{TRS}_{\text{raw}} \text{ within group } g) - 0.5}{n_g} \times 100$$

The -0.5 is Tukey's continuity correction for discrete ranks.

Tie Handling: Assign average rank to tied teams.

3.4.4 Empirical Bayes Shrinkage

Raw PGS from small groups is noisy. Apply shrinkage toward the grand mean (50):

$$\text{PGS}_{\text{shrunk}} = \alpha_g \cdot \text{PGS}_{\text{raw}} + (1 - \alpha_g) \cdot 50$$

where the shrinkage factor is:

$$\alpha_g = \frac{n_g - 1}{n_g - 1 + \kappa}$$

Default: $\kappa = 10$ if estimation is unstable.

Effect of Shrinkage:

- For $n_g = 5$: $\alpha = 4/14 = 0.29$ (71% shrinkage)
- For $n_g = 10$: $\alpha = 9/19 = 0.47$ (53% shrinkage)
- For $n_g = 20$: $\alpha = 19/29 = 0.66$ (34% shrinkage)

3.4.5 Standard Error of PGS

Using the order-statistic approximation:

$$\text{SE}(\text{PGS}_{\text{raw}}) = \frac{50}{\sqrt{n_g}}$$

For shrunken PGS:

$$\text{SE}(\text{PGS}_{\text{shrunk}}) = \alpha_g \cdot \text{SE}(\text{PGS}_{\text{raw}})$$

3.5 Composite CHS Calculation

3.5.1 Aggregation

$$\text{CHS} = w_{\text{CSS}} \cdot \text{CSS} + w_{\text{TRS}} \cdot \text{TRS} + w_{\text{PGS}} \cdot \text{PGS}_{\text{shrunk}}$$

Default weights: $w_{\text{CSS}} = 0.50$, $w_{\text{TRS}} = 0.35$, $w_{\text{PGS}} = 0.15$

3.5.2 Standard Error of CHS

By propagation of uncertainty (assuming component independence):

$$\text{SE}(\text{CHS})_{\text{raw}} = \sqrt{w_{\text{CSS}}^2 \cdot \text{SE}(\text{CSS})^2 + w_{\text{TRS}}^2 \cdot \text{SE}(\text{TRS})^2 + w_{\text{PGS}}^2 \cdot \text{SE}(\text{PGS})^2}$$

Correlation Adjustment:

Components are positively correlated. The independence assumption yields a lower bound. Apply 20% inflation:

$$\mathrm{SE}(\mathrm{CHS}) = 1.2 \cdot \mathrm{SE}(\mathrm{CHS})_{\mathrm{raw}}$$

Note: The 1.2 inflation factor is a provisional conservative estimate based on typical component correlations. It should be empirically validated after 6+ months of production data.

3.5.3 Confidence Interval

Approximate 90% CI:

$$\mathrm{CHS} \pm 1.645 \cdot \mathrm{SE}(\mathrm{CHS})$$

3.6 Indicator and Dimension Structure

The methodology employs a **hierarchical indicator structure** organized into dimensions, each containing multiple indicators that measure related aspects of team health.

3.6.1 Hierarchical Organization

The framework comprises 14 dimensions with approximately 117 indicators:

Dimension	Description	Indicators
1. Invisible Work	Hidden work patterns and visibility gaps	~17
2. Jira as Source of Truth	Information quality and completeness	~18
3. Estimation Practices	Estimation coverage and consistency	~17
4. Issue Type Usage	Consistency in issue type classification	~4
5. Data Freshness	Currency of information	~8
6. Blocker Management	Impediment tracking and resolution	~4
7. Work Hierarchy	Parent-child relationship integrity	~1
8. Sprint Hygiene	Sprint planning and execution discipline	~7
9. Team Collaboration	Communication and cross-functional work	~15
10. Repetitive Work	Duplicate effort detection	~1
11. Automatic Status	Status synchronization accuracy	~5

12. Collaboration Features	Usage of Jira collaboration tools	~5
13. Configuration Efficiency	Workflow and field optimization	~7
14. Backlog Discipline	Backlog health and maintenance	~8

3.6.2 Hierarchical Equal Weighting

The default weighting scheme applies **hierarchical equal weighting**:

1. **Dimension level:** Each dimension receives equal weight: $w_d = \frac{1}{D}$ where D = number of dimensions (14)
2. **Indicator level:** Within each dimension, indicators share the dimension weight equally

For a dimension d with m_d indicators:

$$w_{d,i} = \frac{1}{D} \cdot \frac{1}{m_d} = \frac{1}{D \cdot m_d}$$

This ensures that:

- All dimensions contribute equally to the composite score regardless of indicator count
- A dimension with 17 indicators has the same total weight as one with 4 indicators
- Individual indicators within larger dimensions have proportionally smaller weights

Example calculation:

- Dimension 1 (17 indicators): Each indicator weight = $\frac{1}{14 \times 17} \approx 0.0042$
- Dimension 4 (4 indicators): Each indicator weight = $\frac{1}{14 \times 4} \approx 0.0179$
- Both dimensions contribute $\frac{1}{14} \approx 0.071$ to the total

3.6.3 Indicator Directionality

Each indicator has an assigned directionality:

- **Positive ($d_i = +1$):** Higher values indicate better health (e.g., estimation coverage)
- **Negative ($d_i = -1$):** Lower values indicate better health (e.g., stale work items)

Direction adjustment is applied before aggregation as specified in Section 3.2.1.

3.7 Missing Data Handling

3.7.1 Minimum Coverage Requirements

A team must have valid data for at least 70% of weighted indicators:

$$\sum_{i \in I_t} w_i \geq 0.70$$

Teams below this threshold are excluded from the analysis.

3.7.2 Reweighting Procedure

For partial coverage (≥70% but <100%), reweight available indicators proportionally:

$$w'_i = \frac{w_i}{\sum_{j \in I_t} w_j}$$

3.7.3 Provisional Scores

Teams with < 8 weeks of historical data:

- TRS weight reduces proportionally: $w'_{\text{TRS}} = w_{\text{TRS}} \times (\text{weeks}/8)$
- CSS weight increases to compensate
- Score flagged as "Provisional"

---

4. Threshold Justification

4.1 Interpretation Categories

CHS Range	Category	Description
≥ 70	Excellent Health	Significantly above baseline with strong trajectory
$[55, 70)$	Good Health	Above baseline with positive direction
$[45, 55)$	Average Health	Near baseline norms
$[30, 45)$	Below Average	Room for improvement
< 30	Needs Attention	Significant gaps vs. baseline

4.2 Statistical Basis for Thresholds

With a T-score scale (mean = 50, SD = 15), the thresholds correspond approximately to:

- 70 = 50 + 1.33 SD (approximately 91st percentile under normality)
- 55 = 50 + 0.33 SD (approximately 63rd percentile)
- 45 = 50 - 0.33 SD (approximately 37th percentile)
- 30 = 50 - 1.33 SD (approximately 9th percentile)

The thresholds are symmetric around 50, with the top and bottom bands spanning 20 points (70-100 and 0-30) while the middle bands span 10-15 points, focusing attention

on the tails of the distribution.

4.3 Sensitivity Analysis

Weight sensitivity is mandatory. Report CHS under alternative weight configurations:

Configuration	CSS	TRS	PGS
Balanced (Default)	0.50	0.35	0.15
Snapshot Focus	0.65	0.25	0.10
Growth Focus	0.40	0.45	0.15
Peer Comparison	0.45	0.30	0.25

Sensitivity Criterion: If > 20% of teams change interpretation category across any pair of configurations, report: "Results are sensitive to weight specification. Interpret individual scores with caution."

4.4 Ceiling Effect Handling

Teams with high baseline CHS (≥ 75) have limited room for improvement:

Baseline CHS	Interpretation Guidance
≥ 80	Focus on PGS. Maintaining position indicates "Sustained Excellence"
75-80	Moderate improvement possible. Holding steady is acceptable
< 75	Full interpretation applies

UI Recommendation: When CHS_before ≥ 80 and improvement is minimal, display: "Maintaining Excellence - Your team is operating at a high level and holding steady."

---

5. Scenario Analysis

This section demonstrates methodology behavior through worked examples.

5.1 Complete Worked Example: Single Team CHS Calculation

Setup: Team Alpha has 12 weeks of data across all 14 dimensions. For illustration, we show 4 representative dimensions with varying indicator counts.

Sample Dimension Data (hierarchical equal weighting):

Dimension	Indicators	Dimension Weight	Avg z-score

D3: Estimation Practices	17	1/14 = 0.0714	+1.20
D8: Sprint Hygiene	7	1/14 = 0.0714	+0.85
D9: Team Collaboration	15	1/14 = 0.0714	+1.05
D4: Issue Type Usage	4	1/14 = 0.0714	+0.60
... (10 more dimensions)	...	...	...
Weighted Average			**+0.95**

CSS Calculation:

With hierarchical equal weighting, each dimension contributes equally regardless of indicator count. The dimension-level z-scores are first computed as the average of indicator z-scores within each dimension, then aggregated:

$$\text{CSS}_{\text{raw}} = \frac{1}{14}\sum_{d=1}^{14} z_d = +0.95$$

With $\bar{\rho} = 0.30$ and 14 dimensions:

- Effective variance ≈ 0.35
- Scaling factor: $k = 15/\sqrt{0.35} = 25.4$

CSS = 50 + 25.4 × 0.95 = 74.1

TRS Calculation:

Effect sizes comparing early (weeks 1-6) to recent (weeks 7-12) periods, aggregated hierarchically:

- Average effect size across dimension trajectories: +1.98 SD

TRS = 50 + 10 × 1.98 = 69.8

PGS Calculation:

Team Alpha is in baseline group 3 (8 teams with CSS 30-40 at baseline). Their TRS ranks 7th of 8 within this peer group.

Raw PGS: $(7 - 0.5)/8 \times 100 = 81.25$

With shrinkage ($\kappa = 10$, $\alpha = 0.41$): $$\text{PGS} = 0.41 \times 81.25 + 0.59 \times 50 = 62.9$$

Final CHS:

$$\text{CHS} = 0.50 \times 74.1 + 0.35 \times 69.8 + 0.15 \times 62.9 = 71.0$$

Standard Error: SE = 5.4

90% CI: [62.1, 79.9]

Category: Excellent Health

Key observation: The hierarchical weighting ensures that Team Alpha's strong performance in Estimation Practices (17 indicators) doesn't overshadow their moderate performance in Issue Type Usage (4 indicators)—each dimension contributes ~7.1% to the total.

5.2 Scenario: All Teams Improve Equally

Setup: 50 teams all improve by +15% on all indicators (uniform organizational initiative).

Percentile Analysis:

- All percentiles remain unchanged (invariance problem)
- Stakeholder perception: "No progress visible"

CHS Analysis:

- CSS increases for all teams (measured against fixed baseline)
- TRS shows positive trajectory for all teams
- PGS = 50 for all (no relative difference within groups)
- Overall CHS increases: genuine improvement is visible

Key Insight: CHS successfully detects organizational improvement that pure percentile ranking misses.

5.3 Scenario: High-Performing Team Ceiling Effect

Setup: Team Elite has CSS = 88, operating near maximum.

Analysis:

- Limited room for CSS improvement (max +7 points)
- TRS shows marginal effect size: 53.3
- PGS = 50 (average growth for elite peers)
- CHS = 70.2 (Excellent)

Appropriate Interpretation: "Maintaining Excellence" - sustaining high performance is itself an achievement.

5.4 Scenario: Small Sample (n=15 teams)

Setup: Organization has only 15 teams (below n=20 threshold for PGS).

Model Selection: 2-component model (CSS + TRS only)

Weight Redistribution:

- CSS: 0.59 (was 0.50)
- TRS: 0.41 (was 0.35)
- PGS: 0 (unavailable)

Key Insight: The methodology gracefully degrades when data is insufficient, maintaining valid assessment through available components.

---

6. Implementation Considerations

6.1 Runtime Environment

The methodology is designed for implementation in Atlassian Forge, a serverless JavaScript/TypeScript runtime with specific constraints:

Constraint	Limit	Implication
Execution time	25-55 seconds	No iterative optimization
Memory	Limited	Cannot hold large matrices
Language	JavaScript	No Python/R statistical libraries

6.2 Computational Feasibility

Operation	Feasibility	Notes
CSS calculation	Trivial	Basic arithmetic
TRS calculation	Trivial	Effect sizes, aggregation
PGS grouping	Easy	Sort, group, rank
Empirical Bayes shrinkage	Easy	After variances computed
Analytic SEs	Trivial	Closed-form formulas
Sensitivity analysis (4 configs)	Easy	Run aggregation 4×

6.3 Infeasible Operations and Workarounds

Operation	Issue	Workaround
Quantile regression	No library available	Discrete grouping
Bootstrap (1000 iterations)	Exceeds time limits	Analytic SEs
Complex optimization	No solver libraries	Closed-form estimators

6.4 Caching Strategy

- Cache baseline norms (refresh annually)
- Cache peer group assignments (refresh monthly)

- Cache CHS results (invalidate on new data)

---

7. Limitations and Future Work

7.1 Statistical Limitations

1. **Sample size dependency:** PGS unreliable for n < 20; even with shrinkage, small samples yield imprecise estimates.
2. **Analytic SE approximation:** Less accurate than bootstrap, particularly for PGS near 0 or 100.
3. **Independence assumptions:** Teams treated as independent. Violated if teams share members, projects, or management.
4. **Stationarity assumption:** Baseline norms assumed stable. Organizational changes may invalidate historical comparisons.
5. **Hierarchical equal weighting:** The default weighting scheme treats all dimensions equally, which may not reflect organizational priorities. While this approach is principled (avoiding arbitrary indicator-level weights) and ensures dimensions with more indicators don't dominate, it assumes all 14 dimensions are equally important for measuring team health. Organizations with strong views on dimension prioritization may wish to customize dimension weights; sensitivity analysis should accompany any such customization.

7.2 Inferential Limitations

1. **Association, not causation:** CHS measures health, not the cause of health. Cannot attribute to specific interventions.
2. **Relative components remain:** PGS is still a relative measure. If all teams in a baseline group improve equally, all receive PGS = 50.
3. **Regression to the mean:** Conditioning on baseline partially addresses this. Does not eliminate entirely for extreme baselines.

7.3 Future Work

1. **Empirical validation:** Correlation with external criteria (manager ratings, downstream outcomes); test-retest reliability after 3+ measurement periods.
2. **Weight calibration:** Optimize weights based on predictive validity.
3. **Hierarchical extensions:** Nested team structures (teams within departments); cross-organization comparisons.
4. **Bootstrap implementation:** If runtime constraints relax, implement bootstrap SEs for improved accuracy.

---

8. Conclusion

This paper has presented the Composite Health Score (CHS), a statistical framework for measuring team health in norm-referenced assessment systems while addressing the fundamental invariance problem. When all teams improve uniformly, pure percentile rankings remain unchanged; CHS makes genuine improvement visible through its multi-component design.

The framework combines three complementary measurement approaches: the Current State Score (CSS) measures absolute position against fixed baseline norms, enabling longitudinal comparability without continuous recalibration. The Trajectory Score (TRS) captures improvement momentum using effect-size methodology, reflecting the magnitude of change independent of peer movements. The Peer Growth Score (PGS) contextualizes growth relative to teams with similar starting positions, applying Empirical Bayes shrinkage to stabilize estimates from small groups.

The methodology draws on established statistical foundations from educational measurement (Student Growth Percentiles), psychological measurement (effect sizes), and Bayesian statistics (empirical Bayes shrinkage), adapting these approaches for organizational assessment while remaining computationally feasible in constrained serverless environments.

Key features include transparent uncertainty quantification through analytic standard errors and confidence intervals, mandatory sensitivity analysis to assess weight dependence, and explicit documentation of provisional parameters requiring empirical validation. The complementary Composite Progress Score (CPS) framework extends these principles to measuring change between assessment points.

Practical implications include: organizations can now detect genuine improvement even when all teams improve together; teams starting from challenging positions receive appropriate credit through conditional comparison; and stakeholders receive honest uncertainty estimates rather than false precision. The framework has been designed for production deployment in enterprise software tools.

Limitations include dependence on sample size for PGS reliability, the use of analytic rather than bootstrap standard errors, and provisional weight specifications requiring empirical validation. Future work should focus on empirical calibration of weights and inflation factors, extension to hierarchical team structures, and validation against external criteria.

The CHS methodology provides a principled foundation for organizational health assessment that balances statistical rigor with practical implementation constraints, enabling accurate measurement of team practices and meaningful tracking of improvement initiatives.

References

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42-51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. Guilford Press.

Castellano, K. E., & Ho, A. D. (2013). A practitioner's guide to growth models. *Council of Chief State School Officers*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, 74(1), 68-80. https://doi.org/10.1037/h0029382

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7-29. https://doi.org/10.1177/0956797613504966

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311-319. https://doi.org/10.1080/01621459.1975.10479864

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Sage Publications.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18(8), 519-521. https://doi.org/10.1037/h0049294

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33-50. https://doi.org/10.2307/1913643

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16(4), 421-437. https://doi.org/10.1177/001316445601600401

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381), 47-55. https://doi.org/10.1080/01621459.1983.10477920

OECD. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing. https://doi.org/10.1787/9789264043466-en

Popham, W. J. (1978). *Criterion-referenced measurement*. Prentice-Hall.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50(2), 203-228. https://doi.org/10.1007/BF02294247

Appendix A: Mathematical Derivations

A.1 CSS Variance Formula Derivation

For standardized variables $Z_1, \ldots, Z_m$ with $\text{Var}(Z_i) = 1$ and common pairwise correlation ρ:

$$\text{Var}\left(\sum_{i=1}^{m} w_i Z_i\right) = \sum_{i=1}^{m} w_i^2 \cdot \text{Var}(Z_i) + \sum_{i \neq j} w_i w_j \cdot \text{Cov}(Z_i, Z_j)$$

$$= \sum_{i=1}^{m} w_i^2 + \rho \sum_{i \neq j} w_i w_j$$

$$= \sum_{i=1}^{m} w_i^2 + \rho \left[\left(\sum_{i=1}^{m} w_i \right)^2 - \sum_{i=1}^{m} w_i^2 \right]$$

When $\sum w_i = 1$:

$$= \sum_{i=1}^{m} w_i^2 + \rho \left[1 - \sum_{i=1}^{m} w_i^2 \right] = \sum_{i=1}^{m} w_i^2 (1 - \rho) + \rho$$

A.2 Order-Statistic SE for Percentiles

For order statistics from a uniform distribution on $[0, 1]$, the variance of the k$-th order statistic from n observations is:

$$\text{Var}(U_{(k)}) = \frac{k(n - k + 1)}{(n + 1)^2 (n + 2)}$$

For the median rank ($k \approx n/2$), this simplifies to approximately $\frac{1}{4n}$, giving SE $\approx \frac{1}{2\sqrt{n}}$ on the $[0, 1]$ scale, or $\frac{50}{\sqrt{n}}$ on the $[0, 100]$ scale.

Appendix B: Dimension and Indicator Reference

The CHS methodology employs 14 dimensions containing approximately 117 indicators. This appendix summarizes the dimension structure; complete indicator definitions are maintained in the implementation codebase.

B.1 Dimension Summary

#	Dimension	Purpose	Example Indicators
1	Invisible Work	Detect hidden work and visibility gaps	Throughput variability, stale work items, siloed work
2	Jira as Source of Truth	Assess information completeness	Acceptance criteria coverage, estimates, links

3	Estimation Practices	Evaluate estimation discipline	Story estimation rate, estimate consistency
4	Issue Type Usage	Measure classification consistency	Issue type distribution, volume variability
5	Data Freshness	Track information currency	Stale items, update frequency
6	Blocker Management	Assess impediment handling	Blocker resolution time, blocker documentation
7	Work Hierarchy	Verify structural integrity	Epic linkage coverage
8	Sprint Hygiene	Evaluate sprint discipline	Work carried over, last-day completions
9	Team Collaboration	Measure communication patterns	Comment density, single contributor rate
10	Repetitive Work	Detect duplicate effort	Recreated tickets
11	Automatic Status	Check status synchronization	Stale in-progress work, delayed completions
12	Collaboration Features	Track tool adoption	@mention usage, issue links
13	Configuration Efficiency	Assess workflow optimization	Unused statuses, field load
14	Backlog Discipline	Evaluate backlog health	Zombie items, refinement lag

B.2 Indicator Characteristics

All indicators share common characteristics:

- **Directionality**: Each indicator has assigned polarity (+1 or -1) indicating whether higher values represent better health
- **Winsorization**: Values are bounded at 2nd/98th percentiles to limit outlier influence
- **Report Type**: Each indicator maps to a drill-down report type (issue list, variability chart, distribution, correlation, timeline, or ratio)

B.3 Weight Assignment

Under hierarchical equal weighting:

- Each dimension contributes $\frac{1}{14} \approx 7.1\%$ to the total
- Indicators within a dimension share that dimension's weight equally
- This ensures dimensions with more indicators do not dominate the composite score

Appendix C: Revision History

Version	Date	Changes
CHS 1.0	-	Initial proposal
CHS 1.1	-	Added fixed baseline norms for CSS; fixed aggregate variance scaling; added Empirical Bayes shrinkage to PGS; aligned SE formulas
CHS 1.2	2026-01-26	Corrected CSS variance formula; added ceiling effect guidance; documented SE inflation factors as provisional
Academic Paper 1.0	2026-01-27	Comprehensive documentation for academic review
Academic Paper 1.1	2026-01-27	**Major revision**: (1) Corrected indicator structure to reflect actual 14-dimension hierarchy with ~117 indicators; (2) Updated notation to support hierarchical dimension/indicator structure; (3) Documented hierarchical equal weighting methodology; (4) Revised aggregation formulas for two-stage hierarchical calculation; (5) Updated worked example to demonstrate hierarchical weighting; (6) Rewrote Appendix B with accurate dimension reference

---

Document Version 1.1 | January 2026