Beyond Additive Design: An Empirical Taxonomy of Multimodal STEM Accessibility Systems

Madjid Sadallah
Universite Claude Bernard Lyon 1, CNRS, INSA Lyon, LIRIS, UMR5205
69622 Villeurbanne, France
madjid.sadallah@univ-lyon1.fr

Benoît Encelle
Universite Claude Bernard Lyon 1, CNRS, INSA Lyon, LIRIS, UMR5205
69622 Villeurbanne, France
benoit.encelle@liris.cnrs.fr

Abstract

Multimodal systems combining audio, haptic, and tactile channels are prevalent in STEM accessibility for blind and visually impaired users, yet real-world feedback reports high cognitive load despite technological advances. Through systematic analysis of 66 systems (2015–2025), we identify three architectural regimes: *additive* (channel stacking), *augmentative* (partial coordination), and *integrative* (orchestrated fusion). A five-dimensional scoring framework (ICC=0.92) reveals stark architectural separation, yet conventional performance metrics show no regime variation—a decoupling we term the *Differential Cognitive Yield (DCY)* phenomenon: architecturally distinct systems differ substantially in cognitive cost while yielding similar task performance. We contribute empirical design thresholds for perceptual integration and call for a shift from *interface engineering* to *perceptual integration engineering*.

CCS Concepts

• **Human-centered computing** → **Accessibility**; **Accessibility technologies**; *Multimodal interaction*; • **Computing methodologies** → *Cognitive science*.

Keywords

Accessibility, Multimodality, STEM, Cognitive Engineering, Cognitive Load, Sensory Integration

ACM Reference Format:
Madjid Sadallah and Benoît Encelle. 2026. Beyond Additive Design: An Empirical Taxonomy of Multimodal STEM Accessibility Systems. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3799343

1 Introduction

Multimodality is the dominant paradigm for making STEM content accessible to blind and visually impaired (BVI). By combining auditory, tactile, and haptic channels, researchers aim to create non-visual pathways to spatial reasoning and complex representations [5]. Yet despite technological advances, a persistent gap exists between laboratory validation and real-world adoption: users report excessive cognitive effort, difficulty forming mental models, and reliance on external support even with sophisticated systems [2, 3]. This disconnect suggests the core challenge is not sensor or actuator technology, but rather *how we orchestrate multiple channels into a coherent perceptual experience.*

This challenge connects to established research showing temporal synchronization (<100ms) and semantic congruence are critical for perceptual fusion [19, 23]. When modalities are poorly coordinated, they compete for attention rather than complement each other [24]. Cognitive load theory indicates unintegrated presentation increases extraneous processing [12, 20], particularly problematic for BVI users navigating complex spatial content where poorly integrated information can overwhelm working memory [5, 8].

We identify what we term the *additive fallacy*: the assumption that adding more sensory channels inherently improves access. In practice, many systems *stack* modalities without integrating them, forcing users to manually combine parallel information streams. The HCI accessibility literature currently lacks tools to systematically distinguish between mere channel accumulation and true perceptual integration, leaving designers without clear guidance. Moreover, evaluation practices typically focus on task performance metrics (completion time, accuracy) [7], which may fail to capture the *cognitive cost* of interaction—the mental effort required to integrate information across modalities [6]. A systematic review of 178 BVI evaluations in HCI confirms this pattern: task-completion metrics dominate while cognitive cost measures remain rare [1]. This evaluation gap motivates both our diagnostic framework and the DCY construct we introduce.

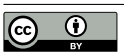
Through empirical analysis of 66 multimodal STEM accessibility systems (2015–2025), we make three contributions. First, we develop a five-dimensional multimodal system scoring framework with high inter-rater reliability (ICC=0.92). Second, this framework identifies three architectural patterns—Additive, Augmentative, and Integrative—with calibrated weights amplifying their separation ($\eta^2 = 0.83$). Third, despite this amplification, conventional performance metrics show *no pattern variation* ($\eta^2 = 0.04$)—a decoupling robust to scoring methodology.

2 Method

Our methodology comprised four sequential phases—corpus compilation, multi-dimensional coding, empirical calibration, and statistical analysis—summarized in Figure 1.

2.1 Corpus and Selection

We compiled 66 peer-reviewed studies (2015–2025) on multimodal STEM accessibility systems, following established systematic review conventions [1, 9]. Sources were drawn from ACM Digital Library, IEEE Xplore, OpenAlex, PubMed, and Web of Science using



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3799343

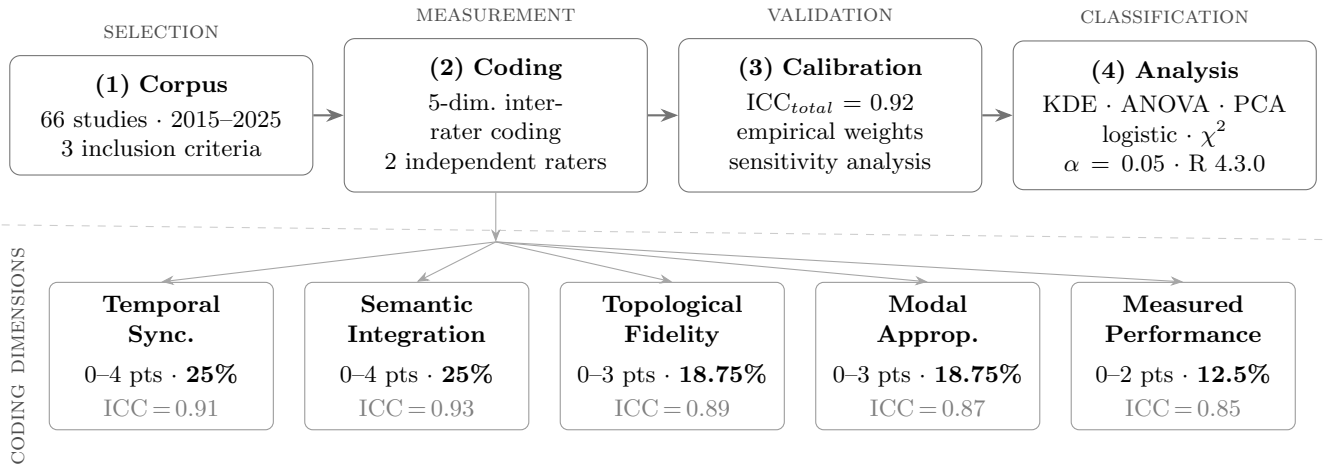


Figure 1: Two-level analytical pipeline. ***Top:*** **four sequential phases with their methodological role (SELECTION, MEASUREMENT, VALIDATION, CLASSIFICATION).** ***Bottom:*** **the five coding dimensions with empirical weights (bold) and inter-rater reliability (ICC).**

search terms combining *multimodal*, *accessibility*, *STEM*, and *non-visual*. Inclusion required: (1) multimodal system (2+ non-visual output modalities) for authentic STEM content; (2) empirical evaluation with BVI participants; (3) sufficient technical detail to code architectural dimensions. The corpus covers mathematics (42.4%), 2D diagrams (33.3%), data visualizations (16.7%), and 3D models (7.6%), with modality combinations of audio+tactile (47.0%), audio+haptic (28.8%), and 3+ modalities (24.2%). This distribution reflects the relative maturity of audio interfaces over haptic and tactile alternatives: richer non-verbal modalities remain underrepresented despite their theoretical advantages for spatial STEM content. Evaluation methodologies included controlled experiments (57.6%), field deployments (27.3%), and mixed methods (15.2%).

2.2 Taxonomy Development

2.2.1 Dimensional Framework. Our five-dimensional framework operationalizes established cognitive constructs following iterative taxonomy development methodology [15], with dimensions scored on their natural scales then weighted to reflect discriminative power revealed through empirical calibration. Two raters coded each dimension independently; disagreements (<12% of cases) were resolved by consensus, with ambiguous cases assigned conservatively to the lower level.

1. Temporal Synchronization (Sync, 0–4 pts, ICC=0.91) captures the temporal coincidence principle [19]. Simultaneous delivery with measured latency ≤100ms scores 4 pts; simultaneous delivery stated or clearly implied without latency measurement scores 3 pts; delivery implied but ambiguous scores 2 pts (conservative); user-triggered presentation scores 1 pt; fixed sequential order scores 0 pts.

2. Semantic Integration (Int, 0–4 pts, ICC=0.93) applies Welch and Warren's unity hypothesis [23]. Synergistic meaning generation unavailable in either channel alone scores 2.5 pts, with a design bonus (+1 pt) when explicit author rationale is provided; complementary information covering different aspects scores 1.5 pts; redundant content scores 1 pt; supplementary detail with no semantic coordination scores 0.5 pt; independent modalities score 0 pts.

3. Topological Fidelity (0–3 pts, ICC=0.89) evaluates spatial structure preservation [5]. Base scores reflect relational encoding: metric spatial relations (distances, proportions, angles) score 2 pts; topological relations (adjacency, containment) score 1.5 pts; hierarchical grouping scores 1 pt; isolated elements score 0.5 pts. Bonuses reward fidelity quality (+1.0/+0.6/+0.3 for full/partial/basic preservation) and navigation support (+0.5–1.0 for landmarks, directional cues, or overview mode).

4. Modal Appropriateness (0–3 pts, ICC = 0.87) evaluates the adequacy of modality selection with respect to content, grounded in dual-coding theory [17]. An explicit alignment between modality and affordances yields 1.5 pts; a design employing a rich non-verbal modality (e.g., haptic, sonification, or tactile) as the primary communication channel yields 1 pt; and a speech-dominated design yields 0.5 pt. Additional richness bonuses are granted as follows: +1.5 pts for ≥ 3 modalities including both tactile and speech; +1.0 pt for ≥ 3 distinct modalities; and +0.5 pt for 2 modalities including sonification.

5. Measured Performance (0–2 pts, ICC=0.85) captures reported empirical outcomes including task accuracy, completion time, cognitive load, knowledge transfer, and retention [20]. Two or more outcome types with significant positive results score 2 pts; a single outcome type with significant results, or multiple outcomes with mixed results, score 1 pt; no user study, no significant results, or informal evaluation scores 0 pts. This rubric intentionally prioritizes breadth and significance of reported outcomes over effect size magnitude, a trade-off appropriate for cross-study comparison but acknowledged as a limit on sensitivity.

2.2.2 Empirical Calibration. Exploratory application to 20 studies (30% of corpus) revealed high Sync-Int correlation ($r = 0.72$), with architectural dimensions explaining 65–70% of variance versus Performance's 3–6%. These observed patterns motivated empirical

recalibration of dimension weights. The final system (Table 1) totals 16 points, with weights reflecting both discriminative power and theoretical orchestration primacy [4, 19, 22, 23]. Sensitivity analysis with equal weights confirms robustness: Performance shows identical regime variation ($\eta^2 = 0.004$ both schemes, $p > 0.87$), validating that the architecture-performance decoupling transcends scoring methodology.

2.3 Reliability and Analysis

Intraclass Correlation Coefficients indicated reliable measurement: Synchronization (0.91), Integration (0.93), Topology (0.89), Modal Appropriateness (0.87), Performance (0.85), with total score reliability of 0.92 [10]. Statistical analyses included: distribution analysis using kernel density estimation to identify natural breaks; one-way ANOVA with Tukey post-hoc tests and effect size calculations (η^2, Cohen's d); Pearson correlation analysis; multinomial logistic regression to establish regime classification thresholds; Principal Component Analysis with varimax rotation; and temporal trend analysis via chi-square tests. All analyses used R 4.3.0 with $\alpha = 0.05$, reporting exact p-values and effect sizes.

3 Results

3.1 Three Architectural Patterns

Score distribution reveals a trimodal structure with natural breaks at 6.0 and 10.5 points, partitioning the corpus into three distinct architectural regimes (Table 2). The regimes are principally distinguished by the co-occurrence of Synchronization and Integration scores: Additive systems exhibit absent temporal coordination (Sync=0.78/4) alongside weak semantic alignment (Int=1.06/4), Augmentative systems show partial progress on both axes, while Integrative systems achieve near-ceiling orchestration (Sync=3.76/4, Int=3.38/4). The regimes exhibit complete statistical separation: 95% confidence intervals are non-overlapping across all three groups, confirming that the breaks reflect genuine discontinuities rather than arbitrary thresholds.

One-way ANOVA confirms that calibrated weights successfully amplify these architectural differences: $F(2, 63) = 152.32$, $p < 0.001$, $\eta^2 = 0.83$ (95% CI [0.75, 0.88]), explaining 83% of total variance. Post-hoc comparisons (Tukey HSD) yield large effect sizes at every transition: Additive vs Augmentative (Cohen's $d = 2.91$ [2.12, 3.70]), Augmentative vs Integrative ($d = 2.89$ [2.08, 3.70]), and Additive vs Integrative ($d = 6.11$ [4.87, 7.35]). The magnitude of the Additive-to-Integrative gap—more than six standard deviations—indicates that these are qualitatively distinct design paradigms, not points on a continuum.

3.2 The Performance-Architecture Paradox

Dimensional analysis reveals systematic architectural evolution (Table 2). Synchronization and Integration exhibit substantial progression: +382% and +219% from Additive to Integrative. Integration shows the largest effect size ($\eta^2 = 0.69$, 95% CI [0.55, 0.77]), explaining 69% of variance despite Synchronization's greater relative improvement (from very low baseline 0.78/4). This confirms semantic coherence as the primary architectural differentiator. Strong Sync-Int correlation ($r = 0.72$, $p < 0.001$, 95% CI [0.58, 0.82]) suggests they function synergistically as an *orchestration construct*.

Yet despite weighting Sync+Int at 50% to reflect theoretical orchestration primacy, Performance shows no significant regime variation: $F(2, 63) = 1.30$, $p = 0.279$, $\eta^2 = 0.04$ (95% CI [0.00, 0.14]). This decoupling proves robust to scoring methodology: equal-weight analysis yields identical results ($\eta^2 = 0.004$, $p = 0.874$), confirming the phenomenon transcends our calibration choices. Performance progression is merely +23% (0.64→0.79/2), compared to 160–382% increases in architectural dimensions. Values remain remarkably low (0.64–0.79 out of 2 possible points), suggesting even best-orchestrated systems struggle to demonstrate performance advantages within current evaluation protocols.

We acknowledge that the coarse 0–2 Performance scale may reduce sensitivity to fine-grained differences. However, this limitation would, if anything, attenuate observable differences—yet the decoupling holds under equal-weight analysis ($\eta^2 = 0.004$, $p = 0.874$), confirming it reflects genuine evaluation inadequacy rather than a measurement artifact. The consistently low scores across all three regimes further indicate that current protocols fail to differentiate architecturally distinct systems regardless of scale granularity [1].

Correlation analysis (Table 3) uncovers systematic relationships. Synchronization and Integration correlate strongly ($r = 0.72$, $p < 0.001$), forming a core orchestration construct. Performance shows consistently weak correlations with all architectural dimensions ($r = 0.19$ to $r = 0.31$), with correlations to Topology ($r = 0.19$, $p = 0.12$) and Modal Appropriateness ($r = 0.24$, $p = 0.052$) failing conventional significance despite N=66—providing additional evidence of architecture-performance decoupling. The correlation structure thus reveals a field where architectural sophistication and measurable outcomes have evolved independently—a structural condition that no recalibration of existing metrics can resolve.

3.3 Empirical Design Thresholds

Multinomial logistic regression identifies precise thresholds governing regime transitions. Moving from Additive to Augmentative requires Sync $\geq$ 1.5/4 AND Int $\geq$ 1.8/4 with 86% classification accuracy (specificity = 89%). Achieving Integrative status demands Sync $\geq$ 3.2/4 AND Int $\geq$ 2.9/4 with 89% accuracy (specificity = 92%). These thresholds are dominated by Sync-Int dimensions (standardized coefficients: $\beta_{\text{sync}} = 0.67$, $p < 0.001$; $\beta_{\text{int}} = 0.71$, $p < 0.001$). The asymmetric requirement for Integrative status (80% synchronization, 72.5% integration) indicates a qualitative leap rather than gradual improvement. Progression from Synchronization to Integration proves more effective than the reverse (area under curve: 0.91 vs 0.78), suggesting reducing temporal latency creates favorable conditions for semantic articulation. Multiple regression reveals 68% of Integration variance is explained by Synchronization and Modal Richness: $R^2 = 0.68$, $F(2, 63) = 66.7$, $p < 0.001$, confirming that semantic integration depends fundamentally on both temporal coordination and thoughtful modality selection [13, 21]. These thresholds and regression findings directly inform the design principles developed in Section 4.

3.4 Principal Component Structure and Temporal Evolution

PCA reduces five dimensions to two components explaining 74.1% variance. We note that applying PCA to ordinal dimensions is a

Table 1: Empirically calibrated scoring system (max 16 points)

Dimension	Operationalization	Max	Wt	ICC
Temporal Synchronization	Synchronous (3) + latency bonus (≤100ms: +1); user-triggered (1); sequential (0)	4	25%	0.91
Semantic Integration	Synergistic (2.5) + design bonus (+1); complementary (1.5); redundant (1); independent (0)	4	25%	0.93
Topological Fidelity	Metric (2); topological (1.5); hierarchical (1); isolated (0.5); +fidelity/navigation bonuses	3	18.75%	0.89
Modal Appropriateness	Multimodal approach (1.5); rich non-visual (1); speech-dominated (0.5); +richness scoring	3	18.75%	0.87
Measured Performance	Task success, cognitive load, transfer, retention (max 2 total)	2	12.5%	0.85
Total		**16**	**100%**	**0.92**

Wt: Weight (percentage of total score) reflecting theoretical orchestration primacy; **ICC:** Intraclass Correlation Coefficient. Performance decoupling validated under equal-weight analysis ($\eta^2 = 0.004$ both schemes).

Table 2: Pattern characterization showing architectural progression and performance paradox

	Regime Overview			Dimensional Architecture					Effect
Regime	**N (%)**	**Total (/16)**	**95% CI**	**Sync (/4)**	**Int (/4)**	**Topo (/3)**	**Modal (/3)**	**Perf (/2)**	**Size (d)**
Additive	25 (37.9%)	4.08	[3.40, 4.76]	0.78 *19.5%*	1.06 *26.5%*	0.84 *28.0%*	0.76 *25.3%*	0.64 *32.0%*	6.11
Augmentative	24 (36.4%)	8.29	[7.68, 8.90]	2.27 *56.8%*	2.67 *66.8%*	1.25 *41.7%*	1.31 *43.7%*	0.79 *39.5%*	2.73
Integrative	17 (25.8%)	12.26	[11.56, 12.96]	3.76 *94.0%*	3.38 *84.5%*	2.18 *72.7%*	2.15 *71.7%*	0.79 *39.5%*	–
Progression A→I				**+382%**	**+219%**	**+160%**	**+183%**	**+23%**	
Statistical significance				$F = 41.36^{***}$	$F = 70.91^{***}$	$F = 20.75^{***}$	$F = 21.63^{***}$	$F = 1.30^{ns}$	
Effect sizes η^2				0.57	0.69	0.40	0.41	0.04	

Dimensions: Sync=Temporal Synchronization, Int=Semantic Integration, Topo=Topological Fidelity, Modal=Modal Appropriateness, Perf=Measured Performance. Italic percentages show utilization relative to dimension maximum. Cohen's d compares to Integrative regime. $^{***}p < .001$, ^{ns} non-significant ($p = .279$). The performance-architecture paradox is visually evident: massive architectural progression (+160–382%) versus minimal performance change (+23%, non-significant).

Table 3: Pearson correlation matrix between the five dimensions (n=66)

	Sync	Int.	Topo.	Modal	Perf.
Synchronization	1.00				
Integration	0.72^{***}	1.00			
Topology	0.62^{***}	0.65^{***}	1.00		
Modal Approp.	0.71^{***}	0.68^{***}	0.54^{***}	1.00	
Performance	0.29^{*}	0.31^{*}	0.19	0.24	1.00

$^{***}p < 0.001$, $^{*}p < 0.05$

known limitation; polychoric PCA would be more strictly appropriate, though the structural patterns we report are consistent with the ordinal correlation structure in Table 3. Component 1 (51.3% variance) represents "Orchestration Quality" (loadings: Sync=0.89, Int=0.91). Component 2 (22.8% variance) represents "Representational Complexity" (loadings: Topo=0.82, Modal=0.76). Performance loads weakly on both axes (0.18, 0.22), confirming its relative independence from architectural dimensions. The three regimes occupy completely disjoint regions with zero overlap in 95% confidence ellipses, providing geometric confirmation of architectural distinctness. These structural findings establish *what* the field has built; temporal analysis reveals *how fast* it is moving.

Analysis of publication trends (2015–2025) reveals significant architectural maturation. Additive approaches decreased from 48% to 28%, Integrative increased from 12% to 31%, while Augmentative remained stable at 35–40% ($\chi^2(4) = 23.4$, $p < 0.001$, Cramer's V=0.30). Evolution follows a two-phase pattern: initial shift to Augmentative (2015–2018), then accelerated Integrative adoption post-2019, coinciding with advances in low-latency audio rendering, transformer-based speech synthesis, and more accessible piezoelectric haptic actuators. This stepwise progression suggests Augmentative as a necessary intermediate step. Three parallel trends accompany: declining speech dominance (65% to 46%), rising explicit multimodal design (15% to 28%), and increasing topological ambition (22% to 38%).

4 Discussion

4.1 Architectural Logics of the Three Regimes

Our calibrated scoring amplifies architectural differences ($\eta^2 = 0.83$), yet Performance resists this amplification ($\eta^2 = 0.04$)—a tension we examine in turn.

The **Additive regime** (37.9%) represents channel accumulation without integration—what Norman terms *featurism* [16]. Poor synchronization (0.78/4) creates integration delays taxing working memory. Each modality imposes processing requirements without

facilitating global integration, forcing users to manually coordinate information streams—a cognitively expensive process that increases extraneous load at the expense of germane processing [20].

The **Augmentative regime** (36.4%) recognizes coordination needs but lacks full integration. Its temporal stability (35–40% across the decade) suggests it functions as a necessary stepping stone rather than final destination. Recent CHI work illustrates this intermediate position: systems combining 3D-printed tactile models with touch-triggered speech labels achieve partial coordination yet still require users to mentally bridge modalities [14, 18].

The **Integrative regime** (25.8%) embodies perceptual integration engineering through temporally coordinated, semantically coherent experiences. That only 25.8% achieve this after a decade indicates absent frameworks rather than technical barriers—the present taxonomy is designed to fill precisely this gap, offering researchers and designers the diagnostic vocabulary needed to target Integrative status intentionally. Integration's largest effect ($\eta^2 = 0.69$) enriches multisensory theory: temporal coincidence is necessary but insufficient without semantic unity [23].

4.2 The Differential Cognitive Yield Phenomenon

Despite dramatic architectural differences (Sync +382%, Int +219%), Performance shows no regime variation ($F = 1.30$, $p = 0.279$, $\eta^2 = 0.04$)—reflecting fundamental evaluation inadequacy rather than measurement noise or genuine regime equivalence. Four mechanisms explain this decoupling. First, only 18.2% of studies incorporated cognitive load measures like NASA-TLX [6]. Second, ceiling effects (87% average success) privilege simple tasks that fail to differentiate systems on complex content. Third, publication bias favors optimal-condition evaluations masking real-world challenges. Fourth, unidimensional metrics ignore learning transfer, retention, and the ability to manage concurrent demands. Furthermore, the 0–2 Performance scale necessarily collapses heterogeneous outcome types—task accuracy, cognitive load, transfer, retention—into a single ordinal index, reducing sensitivity to fine-grained effect sizes and penalizing underpowered evaluations. These methodological limitations do not, however, invalidate the core observation: conventional metrics are designed to capture task completion, not cognitive cost. Users may exhaust working memory via Additive systems while achieving identical accuracy via Integrative systems that preserve capacity for deeper reasoning—a distinction current protocols cannot detect.

We term this the **Differential Cognitive Yield (DCY)** phenomenon: architecturally distinct systems yield similar immediate performance while imposing dramatically different cognitive costs. DCY is a theoretical construct proposed to explain the observed decoupling, not a directly measured outcome—it should be understood as a hypothesis-generating framework rather than a confirmed empirical finding. The low prevalence of cognitive load instrumentation in the corpus (18.2%) underscores the need for an experimental agenda to directly validate whether architecturally superior systems yield measurably lower cognitive cost on complex, authentic tasks. Three potential mechanisms may explain DCY: (1) *Temporal threshold effect*—synchronization $\geq 3.2/4$ may enable automatic perceptual integration versus effortful processing; (2) *Synchronization-integration synergy*—the observed correlation ($r = 0.72$) suggests these dimensions are mutually facilitative; (3) *Representational transformation*—Integrative systems may create unified external representations that reduce mental transformations between modalities. Each mechanism predicts that the cognitive benefits of Integrative design will manifest most clearly on complex, extended tasks where working memory overhead compounds—precisely the conditions absent from the ceiling-effect evaluations documented above. The robustness of this decoupling is empirically confirmed: equal-weight sensitivity analysis yields statistically identical results (Performance $\eta^2 = 0.004$ vs 0.04 calibrated, $p = 0.874$ vs $p = 0.279$), demonstrating DCY is an intrinsic property of current evaluation practices rather than an artifact of our scoring.

4.3 Evidence-Based Design Guidance

Our findings yield two types of actionable guidance. First, empirical thresholds define regime transitions: Augmentative status requires Sync ≥ 1.5 AND Int ≥ 1.8; Integrative status requires Sync ≥ 3.2 AND Int ≥ 2.9 (86–89% classification accuracy). Second, five design principles translate these thresholds into practice:

(1) **Synchronization Primacy**—establish temporal coordination before pursuing semantic integration; synchronization improvement facilitates integration more effectively than the reverse (AUC: 0.91 vs 0.78).
(2) **Semantic Coherence**—design modalities to convey complementary information matched to natural affordances: tactile for spatial exploration, auditory for temporal sequences and alerts.
(3) **Topological Progressivity**—match the complexity of spatial representations to the level of integration already attained; additive systems carrying rich metric content impose unnecessary cognitive load.
(4) **Avoid Additive Traps**—37.9% of surveyed systems remain Additive (Sync=0.78/4, Int=1.06/4), generating cognitive costs that conventional metrics cannot detect.
(5) **Holistic Evaluation**—measure the performance-load- transfer triad rather than isolated task metrics. Incorporating NASA-TLX, dual-task paradigms, transfer tasks, and longitudinal retention studies would expose the costs that ceiling effects currently conceal (87% average success rate) [7, 11].

4.4 Synthesis and Future Directions

Several limitations bound our conclusions. Retrospective coding precludes causal inference: we cannot establish that architectural quality *causes* differential cognitive outcomes, only that the two are systematically decoupled in the literature. Additionally, the high correlation between dimensions ($r = 0.72$) may partly reflect a rater halo effect during coding; if present, such an effect could also inflate ICC estimates, meaning both inter-rater reliability and inter-dimension correlations should be interpreted with some caution. Publication bias may inflate reported Performance scores; corpus coverage reflects English-language peer-reviewed venues and may underrepresent practitioner deployments. Crucially, however, these biases would amplify architecture-performance correlation if better orchestration truly improved outcomes—yet we observe identical decoupling across weighting schemes ($\eta^2 = 0.004$ equal weights,

$\eta^2 = 0.04$ calibrated), strengthening rather than undermining the DCY argument.

The prevalence of Additive approaches (37.9%) indicates a field still prioritizing channel accumulation over perceptual integration. Our thresholds (Sync ≥ 3.2, Int ≥ 2.9) provide measurable targets, but the conceptual shift matters more: from counting channels to orchestrating experiences. Research should experimentally validate DCY, examine long-term learning impacts, and extend this taxonomy beyond STEM. Progress may depend less on how many channels systems have than on how well they work together—a shift from interface engineering toward perceptual integration engineering. Emerging integrative systems that synchronize haptic, tactile, and conversational modalities point toward this future [18].

Acknowledgments

This work is part of the COOBRA project (Coopération entre voyants et déficients visuels braillistes pour apprendre et travailler ensemble), funded by the Agence Nationale de la Recherche (grant ANR-24-SARP-0006) under the Science with and for Society programme (SAPS-RA-RP2).

References

[1] Emeline Brulé, Brianna J. Tomlinson, Oussama Metatla, Christophe Jouffrais, and Marcos Serrano. 2020. Review of Quantitative Empirical Evaluations of Technology for People with Visual Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, Article 290, 13 pages. doi:10.1145/3313831.3376749

[2] Matthew Butler, Leona M. Holloway, Samuel Reinders, Cağatay Goncu, and Kim Marriott. 2021. Technology Developments in Touch-Based Accessible Graphics: A Systematic Review of Research 2010–2020. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 414, 14 pages. doi:10.1145/3411764.3445207

[3] Luiz Felipe da Paixão Silva, Antônio Armando de O. Barbosa, Evelise Roman Corbalan Góis Freire, Paula Christina Figueira Cardoso, Rafael Serapilha Durelli, and André Pimenta Freire. 2018. Content-Based Navigation Within Mathematical Formulae on the Web for Blind Users and Its Impact on Expected User Effort. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*. ACM, New York, NY, USA, 23–32. doi:10.1145/3218585.3218590

[4] Waka Fujisaki, Shinsuke Shimojo, Makio Kashino, and Shin'ya Nishida. 2004. Recalibration of Audiovisual Simultaneity. *Nature Neuroscience* 7, 7 (2004), 773–778. doi:10.1038/nn1268

[5] Nicholas A. Giudice. 2018. Navigating without Vision: Principles of Blind Spatial Cognition. In *Handbook of Behavioral and Cognitive Geography*, Daniel R. Montello (Ed.). Edward Elgar Publishing, Cheltenham, UK, Chapter 15, 260–288.

[6] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Amsterdam, 139–183.

[7] Kasper Hornbæk. 2006. Current Practice in Measuring Usability: Challenges to Usability Studies and Research. *International Journal of Human-Computer Studies* 64, 2 (2006), 79–102. doi:10.1016/j.ijhcs.2005.06.002

[8] Rasha Kadri Ibrahim, Yusraa Ahmed Al Marar, Modhi Salman, Shahed Jehad, Mariam Gaber Hamza, Ahmed Samir Abouelnasr, Sally Mohammed Farghaly Abdelaliem, Shorok Hamed Alahmedi, and Abdelaziz Hendy. 2025. Impact of Multiple Educational Technologies on Well-being: The Mediating Role of Digital Cognitive Load. *BMC Nursing* 24 (2025), 1–12. doi:10.1186/s12912-025-03655-z

[9] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Technical Report TR/SE-0401. Keele University.

[10] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163. doi:10.1016/j.jcm.2016.02.012

[11] Jill H. Larkin and Herbert A. Simon. 1987. Why a Diagram Is (Sometimes) Worth Ten Thousand Words. *Cognitive Science* 11, 1 (1987), 65–100. doi:10.1111/j.1551-6708.1987.tb00863.x

[12] Richard E. Mayer. 2020. *Multimedia Learning* (3rd ed.). Cambridge University Press, Cambridge. doi:10.1017/9781316941355

[13] Richard E. Mayer and Roxana Moreno. 2003. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist* 38, 1 (2003), 43–52. doi:10.1207/S15326985EP3801_6

[14] Ruth Galan Nagassa, Andre Ky Pham, Matthew Butler, Leona Holloway, Kalin Stefanov, Skye de Vent, and Kim Marriott. 2025. Enhancing Tactile Learning: A Co-Designed System for Supporting Speech Interaction with Multi-Part 3D Printed Models by Students who are Blind. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 294, 18 pages. doi:10.1145/3706598.3713706

[15] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European journal of information systems* 22, 3 (2013), 336–359.

[16] Donald A. Norman. 2013. *The Design of Everyday Things* (revised and expanded edition ed.). Basic Books, New York.

[17] Allan Paivio. 1990. *Mental Representations: A Dual Coding Approach*. Oxford University Press, New York.

[18] Samuel Reinders, Matthew Butler, and Kim Marriott. 2025. "It Brought the Model to Life": Exploring the Embodiment of Multimodal I3Ms for People who are Blind or have Low Vision. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 367, 19 pages. doi:10.1145/3706598.3713158

[19] Barry E. Stein and M. Alex Meredith. 1993. *The Merging of the Senses*. MIT Press, Cambridge, MA.

[20] John Sweller, Paul Ayres, and Slava Kalyuga. 2011. *Cognitive Load Theory*. Springer, New York. doi:10.1007/978-1-4419-8126-4

[21] Bruce N. Walker and Michael A. Nees. 2011. Theory of Sonification. In *The Sonification Handbook*, Thomas Hermann, Andy Hunt, and John G. Neuhoff (Eds.). Logos Verlag, Berlin, 9–39.

[22] Mark T. Wallace and Ryan A. Stevenson. 2014. The Construct of the Multisensory Temporal Binding Window and Its Dysregulation in Developmental Disabilities. *Neuropsychologia* 64 (2014), 105–123. doi:10.1016/j.neuropsychologia.2014.08.005

[23] Robert B. Welch. 1999. Meaning, Attention, and the "Unity Assumption" in the Intersensory Bias of Spatial and Temporal Perceptions. In *Cognitive Contributions to the Perception of Spatial and Temporal Events*, Gisa Aschersleben, Talis Bachmann, and Jochen Müsseler (Eds.). Advances in Psychology, Vol. 129. Elsevier, Amsterdam, 371–387.

[24] Christopher D. Wickens. 2002. Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science* 3, 2 (2002), 159–177. doi:10.1080/14639220210123806