

DESIGN RATIONALE • EVIDENCE BASE • OCTOBER 2026

1080 Learner

Design Rationale and Evidence Base

Why a learner-profile tool for K-12 educators gives one page per learner, built on four questions, and what the published research does and does not say about each choice.



01

What we saw

The learner's actual responses, word for word.

02

What it might mean

One reading, always with what else could explain it.

03

Try tomorrow

One practical teaching move for the next lesson.

04

How you'll know

A fresh item, held back, to give in a week or two.

Brandon Robert, Ontario Certified Teacher (registered with the Ontario College of Teachers under his legal name), Milestone Teachers

Research Edition • v1.0
5 October 2026



Not field-tested. 1080 Learner has not been field-tested with real learners. Nothing in this paper is evidence that it works. The research cited here explains why the tool is built the way it is. None of it evaluates 1080 Learner. Research sources reviewed 5 October 2026; access and privacy updated 12 October 2026.

CONTENTS

00	Abstract	2
01	The problem teachers face	3
02	What 1080 Learner does	4
03	Design principles and the research behind each	9
04	What 1080 Learner is not	16
05	Honest limits	17
06	Future evidence	17
07	Access and current terms	19
R	References	20
A	Appendix: evidence table	22



STATUS

1080 Learner has not been field-tested with real learners. Nothing in this paper is evidence that it works. The research cited here explains why the tool is built the way it is. None of it evaluates 1080 Learner. Research sources reviewed 5 October 2026; access and privacy updated 12 October 2026.

Abstract

1080 Learner is a classroom planning tool for K-12 educators in Canada and the United States. A learner completes short math, literacy and thinking tasks, and the educator receives one page per learner answering four questions: What we saw, what it might mean (and what else could explain it), what to try tomorrow, and how you'll know. Items use wrong answers tagged to known error patterns. An evidence engine applies written rules to the responses and records why it chose each item and why it stopped. It calls evidence clear only when at least two responses agree; a single response can give only a tentative reading. This paper sets out each design decision and the research behind it, including its limits and what is still unknown. 1080 Learner has not been field-tested with real learners.

CITATIONS

Each research claim carries an author-year citation and a bracketed claim number, for example [F01]. The companion file REFS.json gives, for every claim number, the source URL, the page or section, and a word-for-word supporting quotation of 25 words or fewer. Every source was opened and read on 5 October 2026. Paywalled studies are cited only “as described in” an open source we did open. In this edition, Appendix A reproduces every entry of REFS.json.

CONFLICT OF INTEREST

The author designed 1080 Learner, and Milestone Teachers sells access to the product. The limits in Section 5 are written with that in mind.

01

The problem teachers face

“A teacher who records only ‘wrong’ loses the most useful part of the response.”

SECTION 1

Two students in the same class each get 6 out of 10 on a short math quiz. They have the same number, but they may need different things tomorrow. Assessment researchers say so directly. An ability estimate alone is of limited use for teaching decisions, because two students with similar estimates can hold completely different ideas (Wylie & Wiliam, 2007) [F09]. The National Research Council (NRC, 2001) makes the same point about counting errors. Three wrong answers on five unrelated subtraction questions show only that some faulty process is at work, not which one [D05].

Right answers can mislead too. Discussing a pair of TIMSS fraction items, Wiliam (2005) argues that many students could answer the easier item correctly by using a wrong rule (smallest denominator, largest fraction) [F13]. Wrong answers, on the other hand, often follow a rule. Falkner, Levi and Carpenter (1999), as described by Knuth et al. (2006), gave students in grades 1 to 6 the sentence $8 + 4 = \square + 5$. Many answered 12 (the left side’s total) or 17 (the total of every number shown). Both answers fit a reading of the equal sign as “announce the result” [MATH-01]. Children often read the sign as meaning “adds up to” or “gets the answer” (Matthews et al., 2012) [MATH-03]. A teacher who records only “wrong” loses the most useful part of the response.

Marks also get in the way. In Butler’s study of 132 students (“year 7” in Wiliam’s account) in 12 classes in four Israeli schools, working on divergent-thinking tasks (not math), as described by Wiliam (2005), comments alone raised later performance, but adding marks to the comments removed the benefit [F04]. Black and Wiliam (1998) concluded that feedback improves learning when it gives specific guidance, preferably without overall marks [F03]. Shute (2007) lists grades that show a student’s standing relative to peers among the features of feedback that tend to impede learning [F05].

Time is the other constraint. In England’s 2016 teacher-workload review, 53% of a sample of respondents to the Workload Challenge survey called the nature, depth and frequency of marking burdensome. The review found very little evidence that extensive written comments improve long-term outcomes (Independent Teacher Workload Review Group, 2016) [F18]. These are English survey data, not Canadian or US data. In the author’s experience the same tension is familiar to teachers in both countries, but we have no comparable data for them.

Policy already asks for something better. Ontario’s *Growing Success* defines assessment for learning as using evidence to decide where learners are, where they need to go and how to get there (Ontario Ministry of Education, 2010) [F15]. It calls for descriptive feedback with specific next steps [F16], and for evidence triangu-

“What teachers most need is not more data but evidence they can use tomorrow morning.”

SECTION 1

02

What 1080 Learner does

lated from observation, conversation and student products [F17]. A US Institute of Education Sciences (IES) practice guide describes a cycle: form a hypothesis from data, change instruction to test it, then collect new data (Hamilton et al., 2009) [MATH-17]. The guide’s own panel rates the evidence for that cycle as low.

Black and Wiliam (1998) put it plainly: assessment becomes formative only when its evidence is actually used to adapt teaching [F01]. In our view, what teachers most need is not more data but evidence they can use tomorrow morning, together with an honest statement of how much weight it can bear.

2.1 A session

A learner works under an alias. The first step, **Tune in**, sets up access before any evidence is collected: text size, read aloud, a display theme (Daylight, Dusk or High contrast) and input by tap or keys. Every setting is printed on the final page, so the adult knows how the tasks were taken. The learner never sees whether an answer was right; every response gets the same reply, “Saved.” “I’m not sure” and “Skip” are full buttons, so nobody is pushed into a guess that would look like evidence. There are no timers.

An **evidence engine**, a set of written rules, works from the learner’s responses. The demo is a fixed-order slice: equal-sign items come first, then fractions. Within a topic it gives a second item because one response cannot settle anything, and it gives a third equal-sign item only when the first two disagree. For each choice it records a plain reason, for example “Chosen because one response cannot settle anything; a second equal-sign item is needed.” Each subject records its own stop reason from a fixed list, such as clear evidence, an item limit, or no suitable item left. When a subject stops at its item limit, the page adds: “This does not show what the learner can or cannot do.” Choosing items by the links between skills is part of the proposed next engine (Section 3.3), not of the demo.

2.2 The page

QUESTION	WHAT THE PAGE GIVES	KIND OF CLAIM
What we saw	The learner’s actual responses and the literal feature they share	Observation
What it might mean + Also possible	One reading, plus at least two other explanations	Interpretation

“The mark shows how much usable evidence there is, not how well the learner did.”

THE PAGE LEGEND

QUESTION	WHAT THE PAGE GIVES	KIND OF CLAIM
Try tomorrow	One practical teaching move	Suggestion
How you'll know	A fresh item, never shown in the session, to give later (math only in the demo; the literacy block asks for one more question on a different passage)	Check

Each kind of claim has its own typeface (monospace for observation, sans serif for interpretation, serif for action) and always a text label. A mark beside each skill shows how much usable evidence stands behind it: filled (clear), half (some evidence), split (mixed), hollow (not enough yet) or dashed (not offered, for example for access reasons). The legend reads: “The mark shows how much usable evidence there is, not how well the learner did.” A “What this is not” footer closes every page. Any single response can be opened in the **Prism**, which splits it into *What they did*, *Also possible* and *Ask next*.

2.3 A worked example: Heron, a sample learner

The public demo has three sample learners. Here is the math path for one of them, Heron. A verification script recomputed every key below across 192 answer-type paths through the demo’s math items (each response counted as correct, a tagged wrong answer, or another answer; paths that use Skip or “I’m not sure” were not enumerated), and it caught a wrong key that we planted to test it.

ITEM	CORRECT ANSWER	HERON’S RESPONSE	TAGGED PATTERN
EQ-1: 7 + 6 = □ + 4	9	13	Total of the left side
EQ-2: 5 + 8 = 6 + □	7	13	Total of the left side
FR-1: 2/3 = □/12	8	11	Added the same amount (9) to top and bottom
FR-2: Which fraction equals 3/4? (6/8, 4/5, 3/8)	6/8	4/5	Added the same amount (1) to top and bottom

Two answers of 13 match the tag for reading “=” as “write the answer next.” (On EQ-1, 17, the total of all three numbers, carries a second tag.) Two consistent responses made the equal-sign evidence clear. The engine therefore skipped EQ-3, a true/false item it gives only when the first two responses disagree, and used the last two of the demo’s four math items on fractions. Both fraction responses add the same amount to top and bottom. One of the two is a three-option choice that can be picked by chance, so the evidence is marked *Some*, not clear. Heron’s page reads, in part:

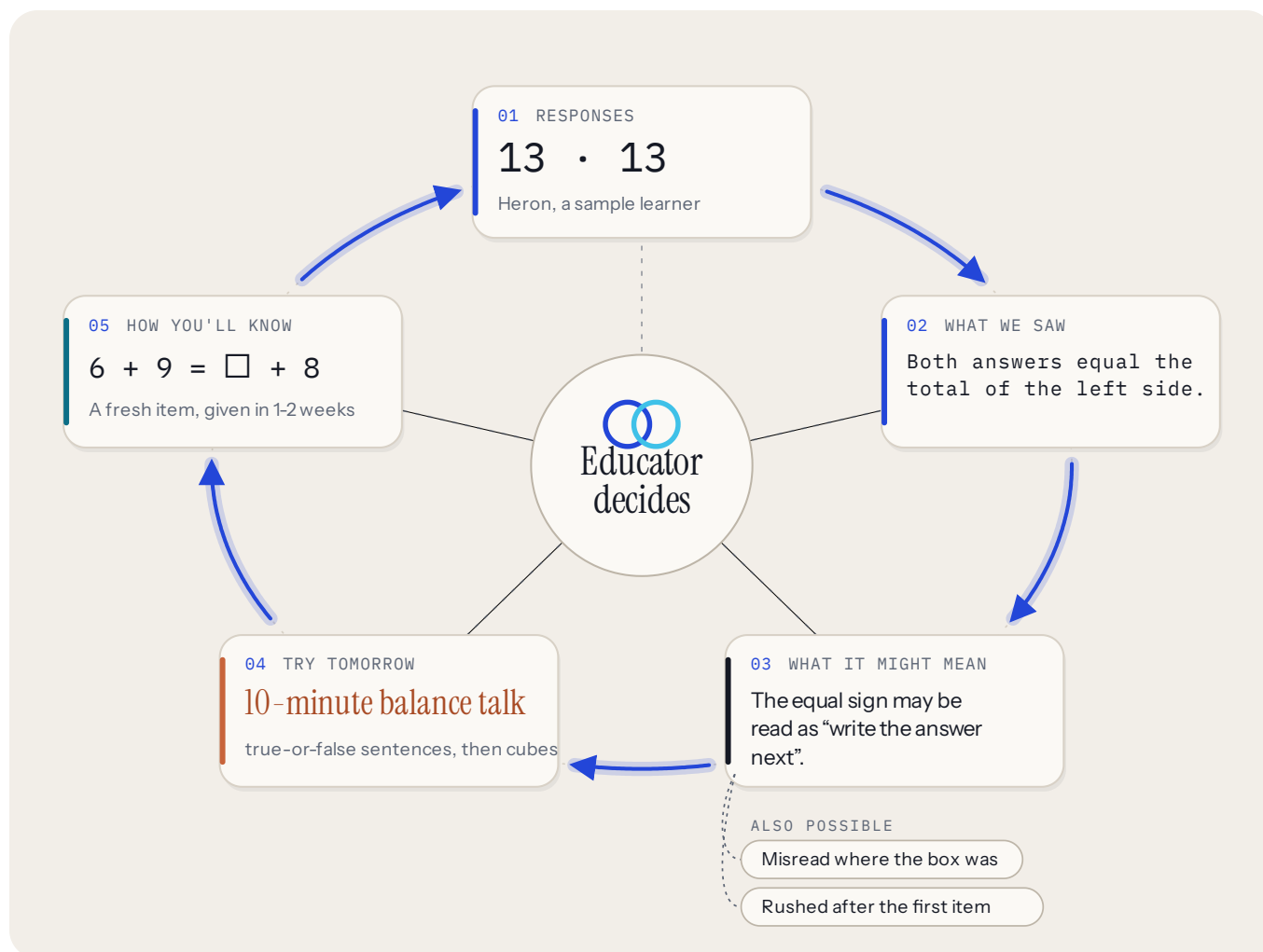


FIGURE 1

The four questions form a loop. The fresh check, not the page, decides whether the move worked.

Loop after the IES data-use cycle (Hamilton et al., 2009; the guide's panel rates its evidence low) [MATH-17]; formative use [F01]. Example values: Heron, a sample learner.

- **What we saw:** “Wrote 13 for $7 + 6 = \square + 4$.” “Wrote 13 for $5 + 8 = 6 + \square$.” “Both answers equal the total of the left side.” “Clear evidence that the equal-sign idea is not yet secure on these items (2 responses).”
- **What it might mean:** “The equal sign may be read as ‘write the answer next.’”
Also possible: “Copied a familiar worksheet habit without reading the right side”; “Misread where the box was”; “Rushed after the first item”.
- **Try tomorrow:** “10-minute balance talk: true-or-false number sentences, then build both sides with cubes.”
- **How you’ll know:** “In 1-2 weeks, give $6 + 9 = \square + 8$. An answer of 7 fits the balance meaning. 15 would be the total of the left side again; 23 would be the total of all three numbers. Different item, so it is not a growth comparison.”

“It never states that Heron ‘has a misconception.’”

SECTION 2.3

- **Equivalent fractions (marked *Some*):** “Wrote 11 for $2/3 = \square/12$.” “For ‘Which fraction is equal to $3/4$?’, chose ‘ $4/5$.’” “Both answers add the same amount to the top and the bottom.” “Some evidence that the equivalent-fraction idea is not yet secure on these items. One more example would confirm it.” Its held-back check is $3/5 = \square/20$: “An answer of 12 fits multiplying the top and the bottom by the same number. 18 would be adding the same amount to both; 4 would be the multiplier written as the answer.”
- **Stop reason (math):** “Stopped because the 4-item limit for this part of the demo was reached.” “This does not show what the learner can or cannot do.”

Notice what the page does not say. It never states that Heron “has a misconception.” The rule for naming a likely misconception requires every plausible setting of the engine’s assumptions to agree, and in the demo it never passes, even on 13 and 13. The page reports the literal feature of the responses and offers the misconception as one reading among several. We will not loosen that rule to make pages sound more confident.

Replay as another learner swaps in a different sample response set: different evidence gives a different page and a different move. Of the 192 answer-type paths, 189 stop at the four-item math limit and carry the caveat. Cedar, the sample learner who answers all four math items correctly, stops because “every idea checked has clear evidence,” and that stop has no caveat. One response is never enough for *clear*, but it can give a tentative *Some evidence* reading. Heron’s single correct literacy answer reads “Some evidence on these items: 1 of 1 correct. One more example would confirm it,” and its move is to wait: “Nothing to change yet. Ask one more question on the same idea before reading anything into this.” The sample learner Delta gave one fraction answer (4 for $2/3 = \square/12$). The page marks it *Some evidence*, not *clear*, but still offers a reading (“The multiplier may have been found and then written as the answer.”) and a reteach move. Cedar’s and Delta’s single literacy answers, both wrong, read “Not enough evidence yet (1 response).” A thinking task (tracking a hidden marker across a grid) is reported only as literal counts, such as “Matched 1 of 1 one-move trial, 1 of 1 two-move trial, 1 of 2 three-move trials.” It comes with a question to ask and a free support idea, and never with a “because,” a comparison or a link.

2.4 The research app (private research app, not yet open)

The research app holds 210 items across 63 sampled skills: 117 math (30 early, 54 middle, 33 secondary) and 93 literacy (30 early, 33 middle, 30 secondary). It also has three thinking tasks, groups of up to 30 learners, and exports as text, JSON, print or a ZIP of separate profiles. The demo items were written for the demo and are not from this bank. The demo runs the evidence rules proposed for the app’s next version (engine v2) with a fixed item order. The research app has not adopted engine v2.



FIGURE 2 One page per learner. Each line says what kind of claim it is.

Wording and values are taken from the public demo’s content files (demo-content v1). Heron is sample. The page is shortened here: the fraction block also carries its own reading, move and fresh check.

03

Design principles and the research behind each

3.1 Wrong answers that carry information

DESIGN Each item is written so that a common wrong rule produces a distinct wrong answer, and that answer is tagged to a written error pattern. For $2/3 = \square/12$, 11 is tagged “added the same amount to top and bottom” and 4 “wrote the scale factor.” On the multiple-choice item, $3/8$ is tagged “changed only the bottom.”

RESEARCH In the US federally funded DIMS project, every multiple-choice item was written so that at least one wrong option maps to a known misconception (Wylie & Wiliam, 2007) [F10]. The same authors argue that items should start from the teacher’s next decision, so that wrong answers show what is not understood and ideally what to do about it (Wylie & Wiliam, 2006) [F14]. The NRC (2001) describes the Facets work of Minstrell and Hunt, in which each answer choice maps back to a way of thinking [D06]. Wang et al. (2020) describe questions whose wrong answers each reveal a common misconception, so educators are not left to guess why [F12]. Sadler’s astronomy items, as described by Wylie and Wiliam (2007), show that a misconception-based wrong option can attract students just below average more than the lowest performers, so right/wrong totals can hide what students believe [F11]; and Wiliam (2005) uses his TIMSS pair to argue that a wrong rule should lead to a wrong answer [F13]. The literature calls these “diagnostic questions.” We say *error-pattern items*, because 1080 Learner is not a diagnosis of any learner.

The patterns come from math research. Equal-sign studies use open equations and true/false structure items [MATH-05], and equations with operations on both sides are generally the harder kind [MATH-04]. 1080 keeps operations on both sides in its session items and its fresh check, so the two stay comparable. For fractions, “whole number bias” describes whole-number knowledge interfering with fractions (Siegler, Thompson & Schneider, 2011, describing Ni & Zhou, 2005) [MATH-10]. The IES fractions guide names treating numerator and denominator as separate whole numbers as a common misconception [MATH-11]. It also describes changing a

denominator without the matching change to the numerator [MATH-12], and calls it critical to find students who hold such misconceptions and discuss them (Siegler et al., 2010) [MATH-13].

LIMIT The IES examples are about fraction addition. Extending them to equivalence, as our “added the same amount” and “changed only the bottom” tags do, is our inference [MATH-11, MATH-12]. Siegler, Thompson and Schneider (2011) also argue that fraction arithmetic errors often reflect confusion about the right strategy, “rather than a consistent whole number bias or other systematic misunderstanding” [MATH-10b]. That is one reason 1080 never treats a single answer as a pattern. Good items of this kind are hard to write: Wang et al. (2020) note that it is challenging to write good items of this kind, in which every wrong answer is a plausible distractor [F12b]. No study has yet confirmed that real learners who give a tagged answer are using the tagged rule.

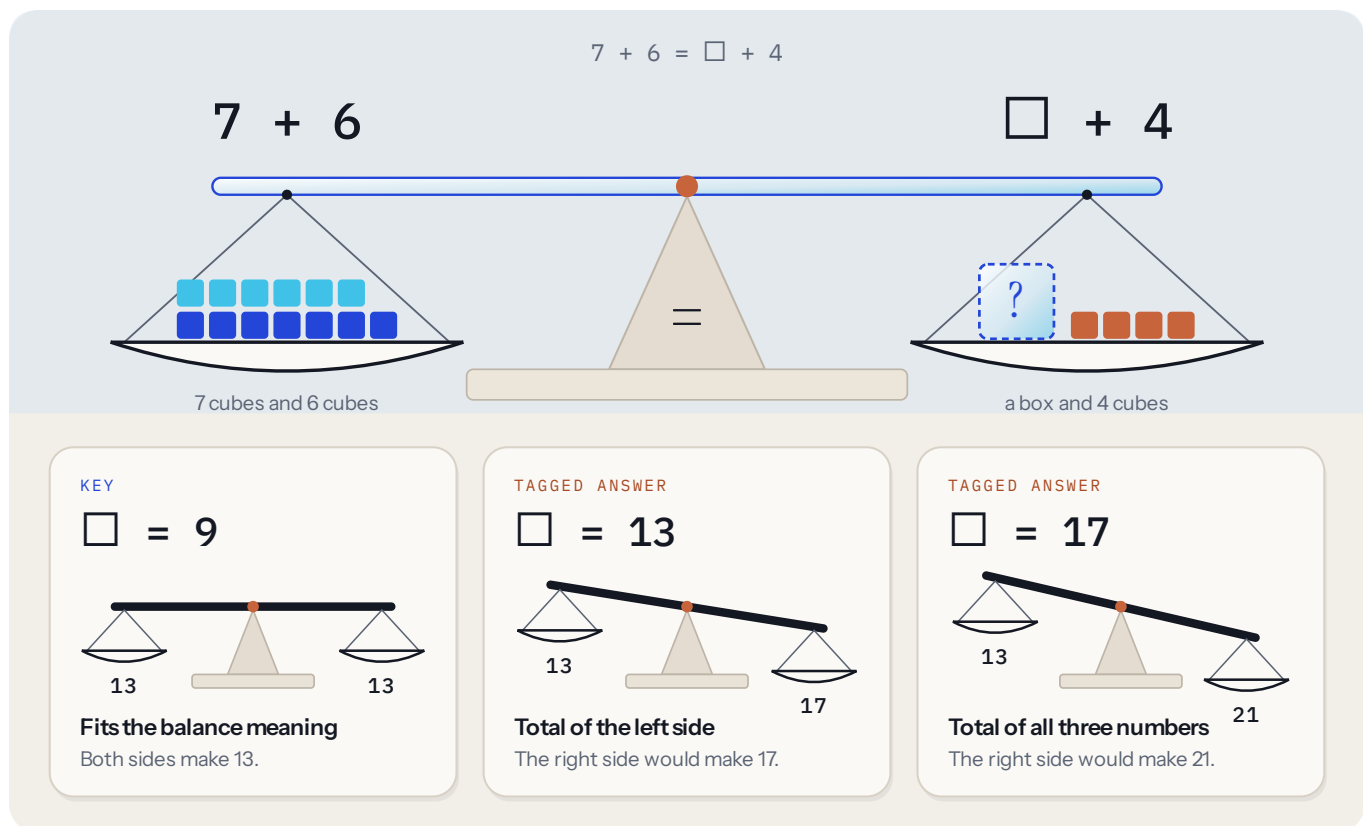


FIGURE 3

Read as a balance, $7 + 6 = \square + 4$ asks for 9. Each tagged wrong answer follows a rule, so it tells you more than “wrong”.

Pattern documented with $8 + 4 = \square + 5$ (Falkner, Levi & Carpenter, 1999, as described in Knuth et al., 2006) [MATH-01]; “adds up to” readings [MATH-03]. Values recomputed from the demo’s verified keys.

3.2 Evidence before labels: “Also possible”

DESIGN No skill is called clear on one response. One response can give only a tentative *Some evidence* reading, which still comes with a reading and a move and says that one more example would confirm it. For *clear*, the engine needs at least

two responses, and it works out each skill's state under eight combinations of unknown settings: a careless-error rate treated as a range (0.08 to 0.30), a guessing rate set by item format, and how consistently a pattern appears. A skill is “clear” only if all eight agree. Every interpretation carries at least two “Also possible” explanations, and the build fails if one is missing. Pattern wording stays literal: the page reports the feature the responses share (for example, “Both answers equal the total of the left side.”). Opening a response in the Prism gives an *Ask next* question.

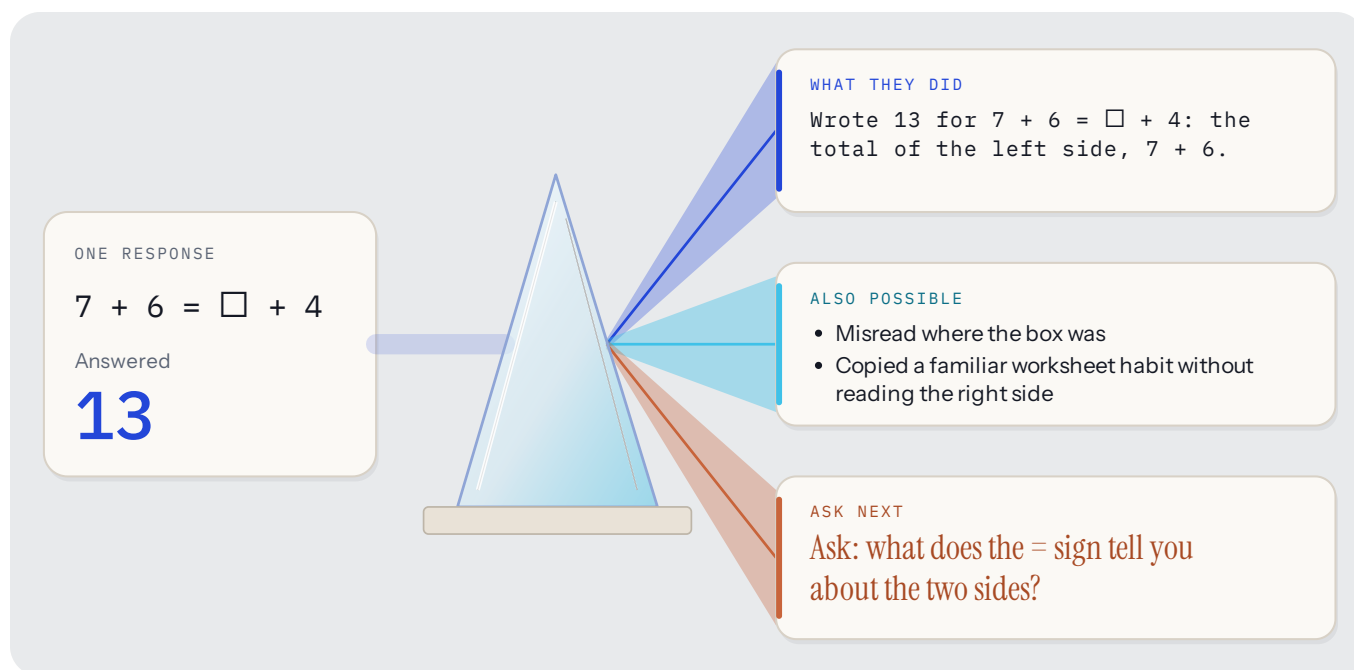


FIGURE 4

One response, three ways of looking at it. The page never names a misconception from a single answer.

A misconception is defined by consistency, not one response [F08]; assessment evidence is open to more than one explanation [D02]. Ask next wording from the demo.

“We will not loosen that rule to make pages sound more confident.”

SECTION 2.3

RESEARCH

The NRC (2001) calls assessment a process of reasoning from evidence [D01]. It notes that assessment information is typically incomplete, inconclusive and open to more than one explanation [D02]. A result is an estimate from a small sample that can vary with format, timing or a student's health that day [D03]. Naming a specific flawed rule takes more than one item [D05], and the model of learning, the tasks and the interpretation method must be designed together [D04]. Wylie and Wiliam (2007) define a misconception by consistency, not by a single wrong answer [F08]. Matthews et al. (2012) warn that either/or results on one item type hide real differences [MATH-06]. The DINA model (Junker & Sijtsma, 2001, as described by Lee, 2017) has a “slip” term (wrong despite mastery) and a “guess” term (right without it) [D11]. 1080's error and guessing ranges make the same allowance.

LIMIT No opened source tests 1080's gate. Applying Matthews et al. to it is our inference.

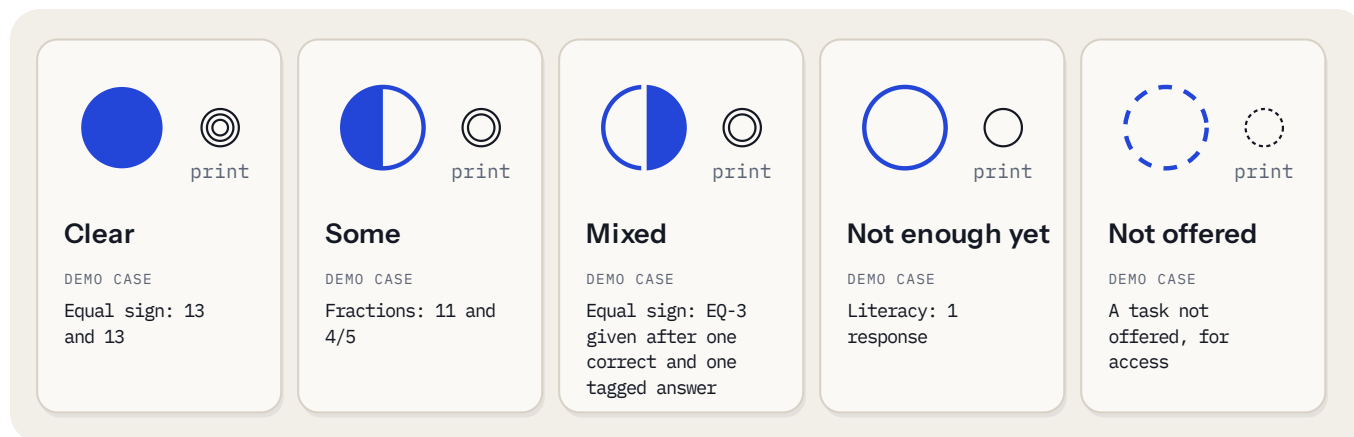


FIGURE 5

How much evidence, not how well. Clear needs at least two agreeing responses, under every plausible setting of careless-error and guessing rates.

[D05], [F08], [D11]. Large mark: screen. Rings: the print form (one to three concentric rings). Thresholds come from internal simulation with synthetic learners only.

3.3 Adaptive questioning and honest stop reasons

DESIGN (proposed: engine v2, not yet in the research app) The proposed engine picks the item most likely to settle the next teaching step and gives priority to skills whose earlier steps look secure. Each item appears at most once. Fresh-check items are held back. No strand takes more than 40% of items, and each subject is capped at 16 items or a time limit. The skill graph chooses items but never sets a reported label. **In the demo**, the item order is fixed (equal-sign items, then fractions), the limits are four math items and one literacy item, fresh checks are held back, and each subject records its own stop reason.

RESEARCH Evidence-centered design separates evidence rules, which read a single response, from the models that combine evidence across tasks (Mislevy, Almond & Lukas, 2004) [D07]. It also names an activity-selection process that decides the next task and when to stop [D08], and an assembly model that sets evidence targets and balance constraints [D09]. 1080 writes its targets and constraints down as the rules above. A simulation study of adaptive tests found that tests that were too short (a mean length under 10 items) did not give good ability estimates for test takers in the low range, and it recommends combining variable stopping rules with a minimum number of items (Babcock & Weiss, 2012) [D10]. The 2014 *Standards* published jointly by AERA, APA and NCME ask that reports say in simple language what was covered, what results mean and how they are meant to be used (American Educational Research Association et al., 2014) [D12]. 1080 prints its evidence counts, stop reason and coverage.

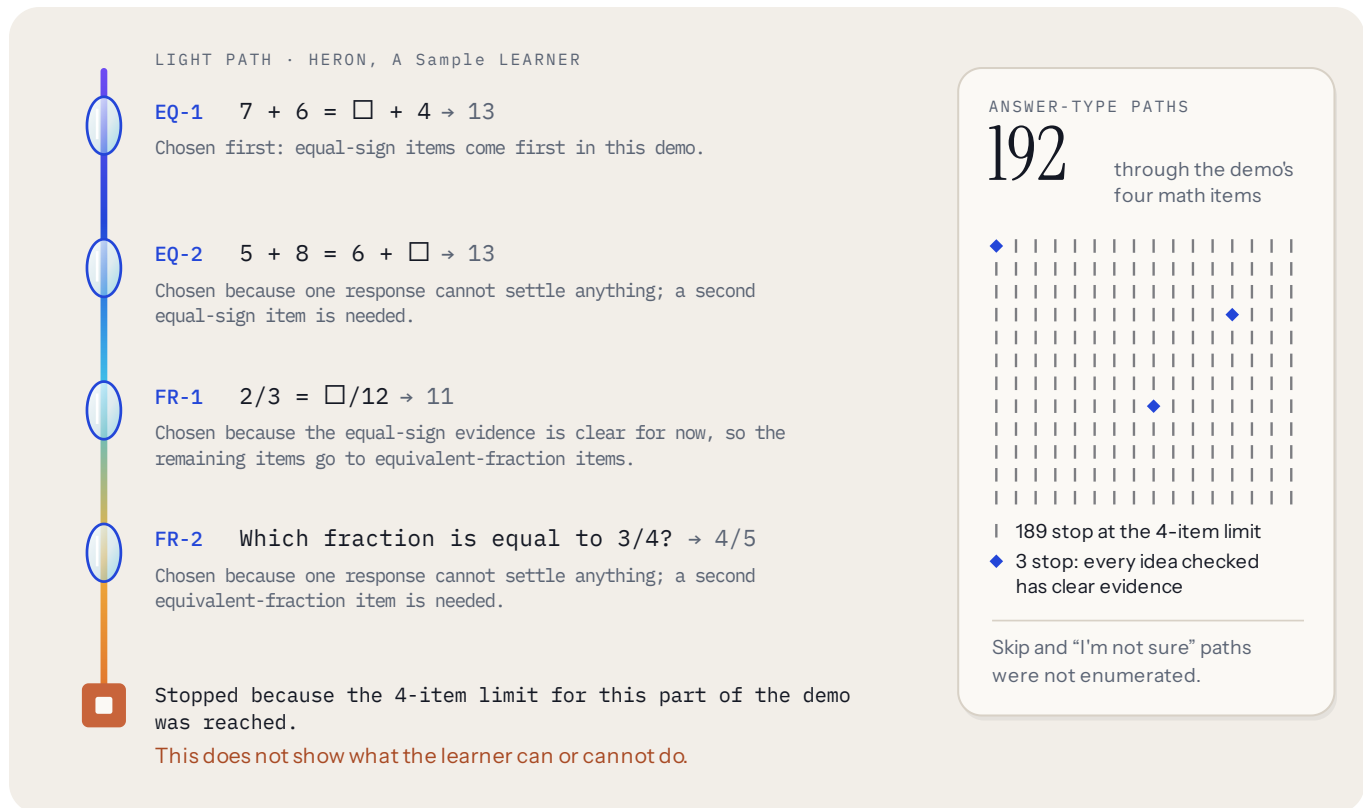


FIGURE 6

Every choice has a stated reason, and every session has one stated stop reason.

Activity selection decides the next task and when to stop [D08]; too-short adaptive tests gave poor estimates in a simulation study [D10]; plain-language reporting [D12]. The inset is a literal count of the demo's enumerated paths.

INTERNAL SIMULATION

(synthetic learners only) We tested the proposed rules on simulated learners (250 to 400 per setting) on an eight-skill test graph. These results describe our design under our own assumptions, not real students. With typed-answer items, wrong confident calls stayed at or below 2.2% for every simulated learner type, in about 12 to 13 items per subject. With four-option multiple choice, a next step was named for only 20 to 27% of simulated learners, which is why typed answers are the authoring target. When we broke the engine on purpose (one response settles a skill, and no guessing is assumed), wrong confident calls rose from 0.9% to 5.0%. The test harness flagged it against a 5% check; the proposed 3% threshold catches it with more margin. Babcock and Weiss studied ability estimates, not per-skill readings, so their work does not support our item counts [D10].

3.4 Next moves and fresh checks

DESIGN

Every skill block ends in a suggested move, shown next to its evidence, and a held-back check in the same format, to be given after a delay. The educator decides whether to use the move, change it or set it aside [D17].

“‘Try tomorrow’ is a practical re-teach, and the fresh check decides whether it worked.”

SECTION 3.4

RESEARCH Formative assessment is defined by use [F01] and works through specific guidance rather than marks [F03]. *Growing Success* asks for specific next steps [F16]. The IES data-use cycle [MATH-17] maps onto 1080’s last three questions. Its evidence is rated low, so we call the loop sensible practice, nothing more. For the equal-sign move, a randomized study (N = 90, mean age 8) found that practice with nontraditional formats such as $17 = 9 + 8$ improved children’s understanding of equivalence more than traditional or no extra practice (McNeil et al., 2011, ERIC abstract) [MATH-08]. The What Works Clearinghouse (2014) rated that study “Meets WWC standards without reservations” for a contrast of 60 students in grades 1 to 3 [MATH-09]. That is one study, and our balance talk itself is untested; the research supports only its non-standard-format ingredient. Equal-sign understanding is *linked to* equation solving in middle school, even after controlling for students’ results on standardized mathematics tests (Knuth et al., 2006) [MATH-02], and to basic algebra in Grades 2 to 6 (Matthews et al., 2012) [MATH-07]. Both findings are correlational. For fractions, the IES guide recommends number lines [MATH-14], and area models and other visual representations for understanding fraction computation [MATH-16], but notes that the number-line evidence comes mainly from whole-number studies [MATH-15]. For timing, an IES guide recommends re-exposure after several weeks to several months (Pashler et al., 2007) [MATH-18]. Our 1 to 2 weeks is shorter, and that guide concerns remembering, not whether a re-check is a sound gauge.

LIMIT As described by Shute (2007), Sleeman et al. (1989) found that simply re-teaching algebra content worked as well as feedback built on detailed error analysis [F07]. Researchers describe feedback findings as inconsistent and highly variable [F06]. Black and Wiliam’s (1998) typical effect sizes of 0.4 to 0.7 come from a 1998 review of formative classroom practice in general and say nothing about 1080 [F02]. We therefore claim no effect. “Try tomorrow” is a practical re-teach, and the fresh check decides whether it worked.

3.5 Families

DESIGN One press turns a page into a **Family Note**: What we saw, Try at home (5 minutes), and What this is not. It targets roughly a Grade 4 reading level, prints on paper, reads on a phone and can be edited by the educator. For now it is in English only. Families receive it through the educator; there are no parent accounts.

RESEARCH US Department of Education guidance recommends that parents’ requests for records held by a provider go through the school or district (Privacy Technical Assistance Center, 2014) [EQ-14]. In a randomized field experiment, weekly one-sentence teacher messages to parents reduced the share of students failing to earn credit, mainly by preventing drop-outs. Messages about what to improve had the largest effects (Kraft & Rogers, 2015, published abstract) [EQ-16]. School email to families at or below the poverty line stayed flat after 2007 (Kraft,

2016) [EQ-17], which is one reason the note works on paper. READY4K! sent parents short, specific activities built on family routines, and its evaluation reports gains in parental involvement (York & Loeb, 2014) [EQ-18].

LIMIT These studies involved US high-school credit recovery and preschool texts. None tested a note like ours.

3.6 Access and Universal Design for Learning

DESIGN Access comes before evidence: text size, themes, read aloud, tap or keys, a tap or key alternative for every drag, no timers and Esc to pause. A task a learner cannot access is recorded as “Not offered: access,” never as wrong. Plain mode removes all art and motion.

RESEARCH CAST’s UDL Guidelines 3.0 (2024) list adjustable text size, spacing and display [EQ-01], name text-to-audio as one of the easiest ways to widen access [EQ-02], and recommend several physical ways to respond [EQ-03]. They also call for informative rather than comparative feedback [EQ-04] and for flexible pace, session length and breaks [EQ-05]. W3C’s Accessibility Guidelines Working Group (2024) recommends WCAG 2.2 as the conformance target [EQ-06]. WCAG 2.2 requires single-pointer alternatives to dragging [EQ-07] and says sign-in should not depend on remembering a password unless an alternative or assisting mechanism is provided (the criterion lists further exceptions) [EQ-08]. We therefore plan for the adult to open each session and the learner to choose an alias, with no learner password.

LIMIT The current build specification aims at WCAG 2.1 AA. We plan to raise the target to 2.2 AA and publish the gap list against it. Voice and switch input are not built yet. Read aloud on a decoding item changes what the item shows; that point is our inference, so the page records when read aloud was on.

3.7 Privacy by design

DESIGN Learners use aliases, and educators never need to enter a learner’s name; adults own the workspace and can export or delete it; each learner has a separate profile; and Milestone Teachers does not sell learner data. The published privacy notice identifies the processors and United States data storage, the retention period and deletion route. The public sample shows status only, with no correctness, grouping or ranking. It uses sample learner responses; it is not a real student record. 1080 makes no predictions.

RESEARCH In October 2025, Canada’s federal, provincial and territorial privacy regulators jointly called for privacy-protective defaults and support for pseudonyms in classroom technology [EQ-09]. They asked vendors to collect only what is necessary and proportionate [EQ-10], named predictions that could shape a child’s educational path as a key risk [EQ-11], and asked for language children can understand [EQ-12]. The Information and Privacy Commissioner of Ontario’s *Digital*

Privacy Charter for Ontario Schools (2024) includes a pledge to make privacy-protective settings the default [EQ-13]. The FTC’s COPPA guidance, which concerns children under 13, counts as personal information “a persistent identifier that can be used to recognize a user over time and across different websites or online services” [EQ-15]. The guidance does not settle whether a classroom alias meets that definition. We therefore treat aliases as personal information and do not rely on them as an exemption. Labels carry their own risk. The NRC (2001) warns that fixed-ability beliefs can lead educators to label students “bright” or “dull” [D15], and teacher expectations have been shown to affect outcomes (Rubie-Davies & Hattie, 2025) [D16]. So 1080 gives no ability labels, and every reading is provisional and paired with a check that can overturn it.

LIMIT

The [privacy notice](#) is published, and learners see a plain-language explanation before opening their session. The notice describes United States storage, retention and deletion. This paper does not independently evaluate legal compliance or the service’s live data handling.

04

What 1080 Learner is not

“We do not claim that using 1080 Learner improves learning.”

SECTION 4

- **Not a diagnosis.** It names no conditions and no causes in a learner. Thinking tasks are reported as what happened in the task, never as a trait.
- **Not a score.** It gives no scores, norms, percentiles, grade equivalents, levels or IQ, and it never compares learners [F05, EQ-04].
- **Not for high-stakes decisions.** It is not for placement, eligibility or special-education planning. The *Standards* say that a decision with major impact on a student should not rest on one test alone [D13], and that where adequate evidence is lacking, users should be told to weigh results accordingly [D14].
- **Not validated.** Validation: none yet. Studies are still needed on consistency, fairness across groups of learners, retest effects and usefulness.
- **Not a replacement for the educator.** A page is one student product. Ontario policy asks for triangulation with observation and conversation [D18]. Opening a response in the Prism gives an *Ask next* question, and educators decide what to do with every suggestion [D17].
- **Not a claim about effect.** We do not claim that using 1080 Learner improves learning.

05

Honest limits

“Nothing in this paper is evidence that it works.”

STATUS

1. **Not field-tested.** No real learner has used 1080 Learner in a pilot. Every engine figure here comes from simulated learners under our assumptions.
2. **Our routines are untested.** The balance talk and number-line reteach are supported only in their ingredients [MATH-08, MATH-14, MATH-15, MATH-16].
3. **The fresh check is not a validated gauge of learning.** Its support is indirect [MATH-17, MATH-18], and our 1 to 2 week delay is shorter than the spacing guide's.
4. **Targeting an error may not beat good re-teaching** [F07], and feedback findings vary widely [F06]. That is why the check decides.
5. **Correlation only.** The equal-sign link to algebra is correlational [MATH-02, MATH-07].
6. **Two tags are our inference.** “Added the same amount” and “changed only the bottom” for equivalence extend addition examples [MATH-11, MATH-12].
7. **Small item sets.** The demo has nine math items (five session items, two held-back fresh checks and two extension items), a practice item, one literacy item and four spatial trials. The app's 210 items sample skills; they are not a curriculum.
8. **Format limits.** If items stay multiple choice, an honest engine names a next step for only about one learner in four (simulation).
9. **A strict naming rule.** The demo never names a likely misconception. We will loosen the rule only if consented response data justify it.
10. **Thinking tasks.** No research cited here supports them. They are reported as literal observations and lead only to free support ideas.
11. **Access gaps.** Voice and switch input are not built. Read aloud on decoding items needs a sourced policy.
12. **English only.** Items and directions are in English. We are not promising a French version.
13. **Data handling has limits.** The current privacy notice explains aliases, United States storage, retention, deletion and the groups the product does not serve.
14. **Evidence from elsewhere.** The workload data come from England [F18], and the family studies come from other age groups [EQ-16, EQ-18].
15. **The author is the developer and seller.**
16. **One response can give a tentative reading.** Only *clear* needs two responses. A *Some evidence* reading can come from a single response, and it still carries a reading and a move.

06

Future evidence

“Counts with their denominators, including null results and results that go against the tool.”

SECTION 6

These are questions for any future research, not results and not a condition of access. No real-learner field study data exist yet.

Participation. Any future study would use a separate invitation and an approved data-handling plan. It would not be designed to show that 1080 Learner improves learning or reported as if it did.

Questions a future study could examine.

1. **Educator-rated usefulness.** After using a page, the educator rates its usefulness for planning on a short fixed scale. They also say whether they used the move, changed it or set it aside, and why.
2. **Agreement with the educator’s own read.** Before opening a page, the educator records their own view of the same skills (secure, not yet or unsure) and any error pattern they have seen in class. We compare the two records. Agreement does not prove either is right, so each disagreement is discussed with the educator and checked against later work.
3. **Change on fresh checks.** Where a move was tried, the educator gives the held-back check after the delay. We record whether the response fits the target idea, repeats the pattern or shows something else. With no comparison group, change cannot be credited to the move, because ordinary teaching, practice and chance also produce change. That question needs a later comparison study.
4. **Time cost.** Learner minutes come from session timestamps. Educator minutes spent reading the page and preparing the move come from brief self-reports.
5. **How the engine works on real responses.** Counts of each stop reason, of “not enough evidence yet” results, of cases where “Also possible” changed what the educator did, and of access settings used. Every reported problem is logged.

Reporting. Any future report should give counts with their denominators, include null and contrary results, keep observation separate from interpretation, contain no individual learner data, and name no school without that school's written permission.

07

Access and current terms

Explore a sample report: <https://milestoneteachers.com/1080/sample-report>. It uses sample data; it is not a real student record.

1080 Learner access. Prices are in Canadian dollars. One-time payment; no automatic renewal.

- **Educator access, CAD 149:** one educator account; 40 pathway starts; 365 days from first adult activation.
- **School access, CAD 990:** one school campus, up to 8 named educator accounts; 240 shared pathway starts; 365 days from first adult activation.

A pathway start is consumed when an educator creates a learner session; resuming it does not consume another start. After payment, find the included code in My purchases, sign in with the educator Google account that will manage the workspace, and redeem the code at milestoneteachers.com/1080/access. Your checkout email and educator Google account can differ. A full refund can be requested within 30 days of purchase at admin@milestoneteachers.ca. A refund ends active access; existing reports remain readable and exportable for 30 days, subject to retention requirements or earlier deletion.

Access terms (scope, refunds and data handling): <https://milestoneteachers.com/legal/1080-access/2026-10-12-r1>

Questions: admin@milestoneteachers.ca

R

References

All sources were opened and read on 5 October 2026. Claim numbers in brackets refer to REFS.json.

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf [D12, D13, D14]
- Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, 7(1). <https://doi.org/10.7333/1212-0101001> [D10]
- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2). Open reprint: <https://kappanonline.org/inside-the-black-box-raising-standards-through-classroom-assessment/> [F01, F02, F03]
- CAST. (2024). *Universal Design for Learning Guidelines version 3.0* (Considerations 1.1, 1.2, 4.1, 7.4 and 8.5). <https://udlguidelines.cast.org/> [EQ-01 to EQ-05]
- Federal, Provincial and Territorial Privacy Commissioners and Ombuds. (2025, October 8). *Protecting the privacy of children and youth through responsible use of educational technologies in the classroom* [Joint resolution]. Office of the Privacy Commissioner of Canada. https://www.priv.gc.ca/en/about-the-opc/what-we-do/provincial-and-territorial-collaboration/joint-resolutions-with-provinces-and-territories/res_20251008_edtech/ [EQ-09 to EQ-12]
- Federal Trade Commission. (n.d.). *Complying with COPPA: Frequently asked questions*. Retrieved October 5, 2026, from <https://www.ftc.gov/business-guidance/resources/complying-coppa-frequently-asked-questions> [EQ-15]
- Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Institute of Education Sciences. <https://web.education.wisc.edu/halverson/wp-content/uploads/sites/33/2012/09/What-Works-DDDM-Report-093009.pdf> [MATH-17]
- Independent Teacher Workload Review Group. (2016). *Eliminating unnecessary workload around marking*. Department for Education (England). <https://assets.publishing.service.gov.uk/media/5a75129f40fb6360e47322f/Eliminating-unnecessary-workload-around-marking.pdf> [F18]
- Information and Privacy Commissioner of Ontario. (2024; the PDF carries no date, and the year comes from its URL path). *Digital privacy charter for Ontario schools*. <https://www.ipc.on.ca/sites/default/files/documents/2024-07-04-Digital%20Privacy%20Charter-8pt5x14.pdf> [EQ-13]
- Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education*, 37(4), 297–312. <https://alibali.psych.wisc.edu/wp-content/uploads/sites/371/2018/02/KnuthStephensMcNeilAlibali2006.pdf> [MATH-01, MATH-02]
- Kraft, M. A. (2016, Spring). The underutilized potential of teacher-parent communication. *Communities & Banking*. Federal Reserve Bank of Boston. <https://www.bostonfed.org/-/media/Documents/cb/The-Underutilized-Potential-of-Teacher-Parent-Communication.pdf> [EQ-17]
- Kraft, M. A., & Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review*, 47, 49–63. <https://doi.org/10.1016/j.econedurev.2015.04.001> (abstract read at <https://ideas.repec.org/a/eee/econedu/v47y2015icp49-63.html>) [EQ-16]
- Lee, S. Y. (2017). *Cognitive diagnosis model: DINA model with independent attributes* [Stan case study]. https://mc-stan.org/learn-stan/case-studies/dina_independent.html [D11]
- Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education*, 43(3), 316–350. Author manuscript: [https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20\(2012\).pdf](https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20(2012).pdf) [MATH-03 to MATH-07]
- McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shiple, H. (2011). Benefits of practicing $4 = 2 + 2$: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development*, 82(5), 1620–1633. <https://doi.org/10.1111/j.1467-8624.2011.01622.x> (ERIC record EJ938672: <https://eric.ed.gov/?id=EJ938672>) [MATH-08]

- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). CRESST, University of California, Los Angeles. <https://cresst.org/wp-content/uploads/R632.pdf> [D07, D08, D09]
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press. <https://doi.org/10.17226/10019> (chapters read at <https://www.nationalacademies.org/read/10019/>) [D01 to D06, D15]
- Ontario Ministry of Education. (2010). *Growing success: Assessment, evaluation, and reporting in Ontario schools* (1st ed., Grades 1 to 12). <https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf> [F15, F16, F17, D17, D18]
- Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007-2004). Institute of Education Sciences. <https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/20072004.pdf> [MATH-18]
- Privacy Technical Assistance Center. (2014). *Protecting student privacy while using online educational services: Requirements and best practices* (PTAC-FAQ-3). U.S. Department of Education. https://studentprivacy.ed.gov/sites/default/files/resource_document/file/Student%20Privacy%20and%20Online%20Educational%20Services%20%28February%202014%29_0.pdf [EQ-14]
- Rubie-Davies, C. M., & Hattie, J. (2025). The powerful impact of teacher expectations: A narrative review. *Journal of the Royal Society of New Zealand*, 55(2), 343–371. <https://doi.org/10.1080/03036758.2024.2393296> [D16]
- Shute, V. J. (2007). *Focus on formative feedback* (ETS Research Report RR-07-11). Educational Testing Service. https://myweb.fsu.edu/vshute/pdf/shute%202007_f.pdf [F05, F06, F07]
- Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., & Wray, J. (2010). *Developing effective fractions instruction for kindergarten through 8th grade* (NCEE 2010-4039). Institute of Education Sciences. https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf [MATH-11 to MATH-16]
- Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, 62, 273–296. <https://www.uni-trier.de/fileadmin/fb1/prof/PSY/PAE/Forschung/SieglerEtAl2011.pdf> [MATH-10, MATH-10b]
- Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Peyton Jones, S., Woodhead, S., & Zhang, C. (2020). *Instructions and guide for diagnostic questions: The NeurIPS 2020 Education Challenge* (arXiv:2007.12061). <https://arxiv.org/abs/2007.12061> (PDF read at <https://arxiv.org/pdf/2007.12061>) [F12, F12b]
- What Works Clearinghouse. (2014). *Review of: Benefits of practicing 4 = 2 + 2* (Study 78771). Institute of Education Sciences. <https://ies.ed.gov/ncee/WWC/Study/78771> [MATH-09]
- Wiliam, D. (2005, January). *Keeping learning on track: Formative assessment and the regulation of learning* [Paper presentation]. 20th Biennial Meeting of the Australian Association of Mathematics Teachers, Sydney, Australia. https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/AAMT%2005%20paper%20%28US%29.pdf [F04, F13]
- World Wide Web Consortium. (2024). *Web Content Accessibility Guidelines (WCAG) 2.2* (W3C Recommendation, 12 December 2024). <https://www.w3.org/TR/WCAG22/> [EQ-06, EQ-07, EQ-08]
- Wylie, C., & Wiliam, D. (2006, April). *Diagnostic questions: Is there value in just one?* [Paper presentation]. Annual meeting of the American Educational Research Association and the National Council on Measurement in Education, San Francisco, CA. https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202006%29.pdf [F14]
- Wylie, C., & Wiliam, D. (2007). *Analyzing diagnostic items: What makes a student response interpretable?* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL. https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202007%29.pdf (the title page reads April 2006; the author site lists it as NCME 2007) [F08, F09, F10, F11]
- York, B. N., & Loeb, S. (2014, rev. 2018). *One step at a time: The effects of an early literacy text messaging program for parents of preschoolers* (NBER Working Paper 20659). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w20659/w20659.pdf [EQ-18]

Studies cited only as described in an opened source (not opened themselves): Butler (1988), as described in Wiliam (2005); Falkner, Levi & Carpenter (1999), as described in Knuth et al. (2006); Junker & Sijtsma (2001), as described in Lee (2017); Ni & Zhou (2005), as described in Siegler et al. (2011); Sadler (1998), as described in Wylie & Wiliam (2007); Sleeman et al. (1989), as described in Shute (2007). Minstrell and Hunt's Facets work is cited as described in NRC (2001).

Internal source: Milestone Teachers. (2026). *Evidence Engine v2: design and simulation with synthetic learners* [Unpublished internal report]. These figures describe our design under our own assumptions and are not research evidence about real learners.

A

Appendix A. Evidence table

Every claim number used in this paper, with the source we opened, where in it the support sits, the access date and a word-for-word quotation of 25 words or fewer. This table reproduces REFS.json. Sources are listed as opened on 5 October 2026; publishers may move or change pages after that date.

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
FORMATIVE ASSESSMENT AND FEEDBACK		
F01	Black & Wiliam (1998) Section "The Argument" (definition paragraph) · opened 2026-10-05 https://kappanonline.org/inside-the-black-box-raising-standards-through-classroom-assessment/	In the formative-assessment research, an assessment counts as formative only when its evidence is actually used to change the next teaching step. “Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs.”
F02	Black & Wiliam (1998) Section "Does Improving Formative Assessment Raise Standards?" · opened 2026-10-05 https://kappanonline.org/inside-the-black-box-raising-standards-through-classroom-assessment/	Black and Wiliam's 1998 review reported typical effect sizes of 0.4 to 0.7 for formative assessment experiments. These are effects of formative classroom practice in general, not of any product, and they say nothing about 1080. “Typical effect sizes of the formative assessment experiments were between 0.4 and 0.7.”
F03	Black & Wiliam (1998) Section "How Can We Improve Formative Assessment?" (subsection "The evolution of effective teaching") · opened 2026-10-05 https://kappanonline.org/inside-the-black-box-raising-standards-through-classroom-assessment/	Feedback improves learning when it gives specific guidance on strengths and weaknesses, preferably without an overall mark. “Feedback has been shown to improve learning when it gives each pupil specific guidance on strengths and weaknesses, preferably without any overall marks.”
F04	Wiliam (2005) p. 6, section "The quality of feedback" · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/AAMT%2005%20paper%20%28US%29.pdf	In Butler's study of 132 students ("year 7" in Wiliam's account) in 12 classes in four Israeli schools, working on divergent-thinking tasks (not math), as described in Wiliam (2005), comments alone raised later performance, while adding marks to the comments removed that benefit. (Wiliam's text dates the study 1998; his reference list gives Butler 1988, British Journal of Educational Psychology 58.) “giving marks alongside the comments completely washed out the beneficial effects of the comments.”
F05	Shute (2007) p. 6 (Section 3, "Feedback and learning") · opened 2026-10-05 https://myweb.fsu.edu/vshute/pdf/shute%202007_f.pdf	A major review of feedback research lists grades or overall scores that show a student's standing relative to peers among the feedback features that tend to impede learning. “Other features of feedback that tend to impede learning include providing grades or overall scores indicating the student's standing relative to peers”
F06	Shute (2007) p. 5 (Section 3, "Feedback and learning") · opened 2026-10-05 https://myweb.fsu.edu/vshute/pdf/shute%202007_f.pdf	Honest limit: researchers who meta-analysed the feedback studies describe the findings as inconsistent and highly variable, so no single feedback design is guaranteed to work. “Researchers who have tackled the tough task of performing meta-analyses on the feedback data use descriptors such as inconsistent, contradictory, and highly variable”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
F07	Shute (2007) p. 11 (feedback complexity; Sleeman et al., 1989, Cognitive Science 13) · opened 2026-10-05 https://myweb.fsu.edu/vshute/pdf/shute%202007_f.pdf	Honest limit: in Sleeman et al. (1989), as described by Shute, remediation built on a model of the student's specific errors was not more effective than simply re-teaching the algebra content. “for procedural skills in algebra, feedback based on just re-teaching content was as effective as feedback based on more expensive error analyses.”
F08	Wylie & Wiliam (2007) p. 3, section "Development of DIMS questions" · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202007%29.pdf	Diagnostic-item researchers define a misconception by consistency: a conception that is reliably activated in a situation, not a single wrong answer (which could be a guess). “a misconception is any conception that is consistently activated in a particular situation, and that results in a student giving an incorrect or incomplete response.”
F09	Wylie & Wiliam (2007) p. 6, section "DIMS Items Meet Traditional Psychometrics" · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202007%29.pdf	A single ability estimate is of limited use for teaching decisions, because two students with the same estimate can hold different ideas; the authors conclude the teacher needs a classification of thinking, not a measurement. “ability estimates by themselves are of limited value in instructional decision-making because two students with similar ability may hold completely different views”
F10	Wylie & Wiliam (2007) p. 3, section "Development of DIMS questions" · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202007%29.pdf	In the DIMS project (US Institute of Education Sciences-funded diagnostic items in math and science), every multiple-choice item was written so that at least one wrong option maps to a known misconception ("distractor-driven" items). “at least one misconception is addressed by every diagnostic item, so that the items are, in a sense distractor-driven (Sadler, 1995).”
F11	Wylie & Wiliam (2007) pp. 4-5 (Figure 1 discussion; Sadler 1998, Journal of Research in Science Teaching 35(3), as described) · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202007%29.pdf	Sadler's (1998) distractor-driven astronomy items, as described in Wylie and Wiliam, showed a misconception-based wrong option attracting students just below average more than the lowest performers, so right/wrong totals can hide what students believe. “for students just below average ability, the probability of a correct response is almost 20 percentage points lower”
F12	Wang, Lamb, Saveliev, ... Barton, Peyton Jones, Woodhead & Zhang (2020) PDF p. 3, section "Background" (the quote is in the PDF, not on the arXiv abstract page) · opened 2026-10-05 https://arxiv.org/pdf/2007.12061	Diagnostic questions are multiple-choice items whose wrong answers are each chosen to reveal a common misconception, so a wrong answer tells the teacher something specific. “If a student gets a diagnostic question wrong, educators are not left to guess why.”
F12b	Wang, Lamb, Saveliev, ... Barton, Peyton Jones, Woodhead & Zhang (2020) PDF p. 4 · opened 2026-10-05 https://arxiv.org/pdf/2007.12061	Wang et al. note that good diagnostic questions, in which every wrong answer is a plausible distractor, are challenging to write. “It is challenging to write good diagnostic questions, where each of the incorrect answers is a plausible distractor.”
F13	Wiliam (2005) pp. 1-2, section "What makes a good question?" · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/AAMT%2005%20paper%20%28US%29.pdf	Discussing a pair of TIMSS fraction items, Wiliam argues that many students could answer the easier one correctly using a wrong rule (smallest denominator means largest fraction), inferring this from the response percentages, so a correct answer alone can hide a misconception. “In this sense, the first item is a much weaker item than the second, because many students can get it right for the wrong reasons.”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
F14	Wylie & Wiliam (2006) p. 3, section "Characteristics of 'Decision-Pull' Questions" · opened 2026-10-05 https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers_files/DIMS%20%28NCME%202006%29.pdf	Diagnostic-item researchers argue items should start from the teacher's next decision ("decision-pull"): wrong answers should show what is not understood and, ideally, what to do about it. "Incorrect responses assist the teacher diagnose what students do not understand, and, ideally, provide ideas about what to do about this."
F15	Ontario Ministry of Education (2010) Chapter 4, p. 31, Table 4.1 (quoting Assessment Reform Group, 2002) · opened 2026-10-05 https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf	Ontario's assessment policy defines assessment for learning as using evidence to decide where learners are, where they need to go and how best to get there. "seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go"
F16	Ontario Ministry of Education (2010) Chapter 4, p. 34, section "Providing Descriptive Feedback" · opened 2026-10-05 https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf	Growing Success describes descriptive feedback as precise information on what the student does well, what needs improvement and specific steps to improve. "providing them with precise information about what they are doing well, what needs improvement, and what specific steps they can take to improve"
F17	Ontario Ministry of Education (2010) Chapter 4, p. 34, section "Eliciting Information about Student Learning" · opened 2026-10-05 https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf	Ontario policy asks teachers to triangulate evidence from observation, conversations and student products rather than rely on one source. "These strategies should be triangulated to include observation, student- teacher conversations, and student products."
F18	Independent Teacher Workload Review Group, Department for Education, England (2016) p. 6, para. 9, section "What is the problem?" · opened 2026-10-05 https://assets.publishing.service.gov.uk/media/5a75129f40f0b6360e47322f/Eliminating-unnecessary-workload-around-marking.pdf	In England's 2016 teacher-workload review, 53% of a sample of respondents to the Workload Challenge survey said the nature, depth and frequency of marking was burdensome; the review (p. 5, para. 5) found very little evidence that extensive written comments improve long-term outcomes. "53% of sample respondents thought that, whilst marking pupils' work is necessary and productive, the excessive nature, depth and frequency of marking was burdensome."

ASSESSMENT DESIGN AND REPORTING

D01	National Research Council (2001) p. 53 (chapter conclusions) · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/4	The National Research Council's 2001 report on assessment calls assessment a process of reasoning from evidence, and it says the results are only estimates of what a person knows and can do. "One such principle is that assessment is always a process of reasoning from evidence."
D02	National Research Council (2001) p. 42 · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/4	The same report says the information an assessment produces is typically incomplete and inconclusive, and can usually be explained in more than one way. "the information produced by an assessment is typically incomplete, inconclusive, and amenable to more than one explanation."
D03	National Research Council (2001) p. 42 · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/4	The report describes an assessment result as an estimate drawn from a small sample of everything a learner knows and can do. It adds that results can vary for reasons unrelated to achievement, such as item format, timing or the student's health that day. "an assessment result is an estimate, based on samples of knowledge and performance"

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
D04	National Research Council (2001) p. 54 (triangle introduced p. 44) · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/4	The NRC 'assessment triangle' says every assessment rests on three linked elements: a model of how students learn (cognition), the tasks that provide evidence (observation) and a method for making sense of that evidence (interpretation). The report says these must be designed together. “These three elements—cognition, observation, and interpretation—must be explicitly connected and designed as a coordinated whole.”
D05	National Research Council (2001) p. 200 · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/9	The NRC report warns that counting wrong answers on unrelated items shows only that something is going wrong, not what. It says naming a specific flawed rule takes more than one item, and a set of items designed to work together. “If a student answers three of five separate subtraction questions incorrectly, one can infer only that the student is using some faulty process(es)”
D06	National Research Council (2001) p. 203 · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/9	The NRC report describes multiple-choice items, from the Facets work of Minstrell and Hunt, in which each answer choice is written to match one particular way of thinking. A wrong answer then carries information instead of only being 'wrong'. “Multiple-choice questions have also been designed to identify facets of student thinking by having each of the answer choices map back to a particular facet.”
D07	Mislevy, Almond & Lukas (2004 CRESST release; ETS lists the same paper as RR-03-16, 2003) p. 10 · opened 2026-10-05 https://cresst.org/wp-content/uploads/R632.pdf	Evidence-centered design (ECD) separates two steps. 'Evidence rules' pick out what matters in a single response. 'Measurement models' then combine that evidence across tasks. “evidence rules concern the identification and summary of evidence within tasks, in terms of observable variables.”
D08	Mislevy, Almond & Lukas (2004 CRESST release; ETS lists the same paper as RR-03-16, 2003) p. 16 · opened 2026-10-05 https://cresst.org/wp-content/uploads/R632.pdf	In ECD's four-process delivery model, an Activity Selection process decides which task comes next and when the assessment stops. In adaptive assessments it does so using the current state of what is known about the student. “This process is responsible for deciding what the next task should be and when to stop the assessment.”
D09	Mislevy, Almond & Lukas (2004 CRESST release; ETS lists the same paper as RR-03-16, 2003) pp. 12-13 · opened 2026-10-05 https://cresst.org/wp-content/uploads/R632.pdf	ECD's Assembly Model sets two things: targets for how accurately each skill must be measured, and constraints that keep the set of tasks balanced across the domain. “Targets describe how accurately each student model variable must be measured, and constraints describe how tasks must be balanced”
D10	Babcock & Weiss (2012) printed p. 14 (PDF p. 15), Discussion (result); recommendation on printed p. 15 (PDF p. 16) · opened 2026-10-05 https://jcatpub.net/index.php/jcat/article/download/16/3	A simulation study of stopping rules for computerized adaptive tests found that tests that were too short (mean length under 10 items) did not give good ability estimates in the low range of ability. The authors recommend combining variable stopping rules with a minimum number of items. “CATs that were too short (e.g., mean length less than 10 items) did not give good θ estimates in low ranges of θ ”
D11	Lee, S. Y. (2017), describing Junker & Sijtsma (2001) Section 1.1 'Overview' · opened 2026-10-05 https://mc-stan.org/learn-stan/case-studies/dina_independent.html	The DINA cognitive diagnosis model (Junker & Sijtsma, 2001, as described in an open Stan case study) includes a 'slip' probability (a wrong answer although the skill is mastered) and a 'guess' probability (a right answer although it is not). “The slip parameter s_i is the probability that respondent j responds incorrectly to item i although he or she has mastered all required attributes.”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
D12	AERA, APA & NCME (2014) p. 119 · opened 2026-10-05 https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf	The 2014 Standards (AERA, APA, NCME) say a reported result should explain, in plain language, what the assessment covers, what the results mean, how precise they are and how they are meant to be used. “describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used”
D13	AERA, APA & NCME (2014) p. 198 · opened 2026-10-05 https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf	The Standards say that in schools, a decision or description with a major impact on a student should not rest on one test alone. It should also draw on other relevant information, such as classroom observation and coursework. “a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test”
D14	AERA, APA & NCME (2014) p. 199 · opened 2026-10-05 https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf	For results used in placement-type decisions, the Standards say that where adequate empirical evidence is not available, users should be told to weigh the results accordingly. “When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information”
D15	National Research Council (2001) p. 235 · opened 2026-10-05 https://www.nationalacademies.org/read/10019/chapter/10	The NRC report warns that believing ability is fixed can lead educators to label students, openly or quietly, as 'bright' or 'dull', and that tracking can then reinforce those labels. “Such beliefs on the part of educators can lead both to the labeling—overtly or covertly—of students as 'bright' or 'dull'”
D16	Rubie-Davies & Hattie (2024 online; 2025 issue) p. 343 · opened 2026-10-05 https://minerva-access.unimelb.edu.au/server/api/core/bitstreams/a9dd796b-2a37-4266-92dd-04564aa51285/content	A 2025 open-access narrative review notes that teacher expectations have been shown to affect student outcomes. It reports large self-fulfilling prophecy effects for teachers with strong stereotypical bias. “Teacher expectations have been shown to affect student outcomes.”
D17	Ontario Ministry of Education (2010) p. 8 (Ch. 1 'Fundamental Principles') · opened 2026-10-05 https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf	Ontario's assessment policy puts teachers' professional judgement at the centre of assessment, evaluation and reporting. “Teachers' professional judgements are at the heart of effective assessment, evaluation, and reporting of student achievement.”
D18	Ontario Ministry of Education (2010) p. 34 · opened 2026-10-05 https://www.edu.gov.on.ca/eng/policyfunding/growSuccess.pdf	Ontario policy says assessment for learning should be triangulated across observation, student-teacher conversations and student products. “These strategies should be triangulated to include observation, student-teacher conversations, and student products.”

MATH ERROR PATTERNS AND TEACHING

MATH-01	Knuth, Stephens, McNeil & Alibali 2006 (describing Falkner, Levi & Carpenter 1999) p. 298 (PDF p. 3) · opened 2026-10-05 https://alibali.psych.wisc.edu/wp-content/uploads/sites/371/2018/02/KnuthStephensMcNeilAlibali2006.pdf	In Falkner, Levi & Carpenter (1999), as described in Knuth et al. (2006), students in grades 1-6 solving $8 + 4 = \boxed{} + 5$ often answered 12 or 17: the left-side total, or the sum of every number shown. Both answers fit reading the equal sign as 'announce the result'. “many students provided answers of 12, 17, or 12 and 17 - answers that are consistent with an understanding of the equal sign as announcing a result”
----------------	---	--

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
MATH-02	Knuth, Stephens, McNeil & Alibali 2006 p. 297, abstract (PDF p. 2); the analysis controlling for test scores is on p. 305 · opened 2026-10-05 https://alibali.psych.wisc.edu/wp-content/uploads/sites/371/2018/02/KnuthStephensMcNeilAlibali2006.pdf	Among middle school students, a relational understanding of the equal sign was associated with success at solving equations measured at the same time (a concurrent, not a later, link), and the association held after controlling for standardized mathematics test scores. The finding is correlational and does not show cause. “their understanding of the equal sign is associated with performance on equation-solving items. Moreover, the latter finding holds even when controlling for mathematics ability”
MATH-03	Matthews, Rittle-Johnson, McEldoon & Taylor 2012 manuscript p. 222 (PDF p. 3) · opened 2026-10-05 https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20(2012).pdf	Many children read the equal sign as an operator signal, meaning ‘adds up to’ or ‘gets the answer’, and do not read it as a statement that both sides are equal. “Children often interpret the equal sign as an operator signal that means ‘adds up to’ or ‘gets the answer’”
MATH-04	Matthews, Rittle-Johnson, McEldoon & Taylor 2012 manuscript p. 234 (PDF p. 15) · opened 2026-10-05 https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20(2012).pdf	Item format changes difficulty. In a Grades 2-6 sample (N = 224), equations with operations on both sides of the equal sign were harder than equations with operations on one side only. “equations with operations on both sides were generally more difficult than equations with operations only on a single side”
MATH-05	Matthews, Rittle-Johnson, McEldoon & Taylor 2012 manuscript p. 223 (PDF p. 4) · opened 2026-10-05 https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20(2012).pdf	Equal-sign research uses several distinct item types: open equations ($8 + 4 = [\text{box}] + 5$), equation-structure items (judging whether a number sentence is true or false) and requests to define the equal sign. “(b) equation-structure items, such as deciding whether $3 + 5 = 5 + 3$ is true or false”
MATH-06	Matthews, Rittle-Johnson, McEldoon & Taylor 2012 manuscript p. 226 (PDF p. 7) · opened 2026-10-05 https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20(2012).pdf	Classifying children as passing or failing one type of item hides real differences in what they know. Matthews et al. argue for drawing on several item types. “Either-or distinctions hide the within-group variability in the binary groups that result.”
MATH-07	Matthews, Rittle-Johnson, McEldoon & Taylor 2012 manuscript p. 243 (PDF p. 24) · opened 2026-10-05 https://faculty.weber.edu/eamsel/Research%20Groups/Math%20Research/Mathews%20et%20al.%20(2012).pdf	In Grades 2-6, children with more advanced equal-sign knowledge were more likely to solve basic algebra items, including items with letters as variables (e.g., $c + c + 4 = 16$). “Children with more advanced knowledge of the equal sign were more likely to solve both types of algebraic items correctly.”
MATH-08	McNeil, Fyfe, Petersen, Dunwiddie & Brletic-Shipley 2011 (ERIC abstract) ERIC record, abstract · opened 2026-10-05 https://eric.ed.gov/?id=EJ938672	In a randomized study (M age 8;0, N = 90), children who practised addition written in a nontraditional format (e.g., $17 = 9 + 8$) developed a better understanding of mathematical equivalence than children who had traditional-format practice or no extra practice. “Children developed a better understanding of mathematical equivalence after receiving nontraditional practice than after receiving traditional practice or no extra practice.”
MATH-09	What Works Clearinghouse (IES), review dated April 2014, of McNeil et al. 2011 Review Details, April 2014 review (Nontraditional Addition Problem Formats) · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Study/78771	The What Works Clearinghouse reviewed the McNeil et al. (2011) nontraditional-format study as a randomized controlled trial and gave it its highest rating. It reported statistically significant positive findings on defining the equal sign and on equation solving (60 students, grades 1-3, in the reviewed contrast). “Rating: Meets WWC standards without reservations”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
MATH-10	Siegler, Thompson & Schneider 2011 (describing Ni & Zhou 2005) p. 274 (PDF p. 2) · opened 2026-10-05 https://www.uni-trier.de/fileadmin/fb1/prof/PSY/PAE/Forschung/SieglerEtAl2011.pdf	Conceptual-change accounts of fraction learning, including Ni & Zhou (2005), emphasize a 'whole number bias': knowledge that works for whole numbers interferes when it is applied to fractions. “emphasizing how the ‘whole number bias’ interferes with fractions learning”
MATH-10b	Siegler, Thompson & Schneider 2011 p. 292 (PDF p. 20) · opened 2026-10-05 https://www.uni-trier.de/fileadmin/fb1/prof/PSY/PAE/Forschung/SieglerEtAl2011.pdf	The same paper argues that fraction arithmetic errors often reflect confusion about the right strategy, together with a lack of constraints on answer size, rather than a consistent whole number bias or other systematic misunderstanding. “rather than a consistent whole number bias or other systematic misunderstanding”
MATH-11	Siegler, Carpenter, Fennell, Geary, Lewis, Okamoto, Thompson & Wray 2010 p. 32 (PDF p. 38), Recommendation 3 · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf	The IES fractions practice guide names treating numerators and denominators as separate whole numbers as a common misconception. Its example is adding the tops and the bottoms ($2/4 + 5/4 = 7/8$), which it explains as misapplied whole-number addition. “Students who err in this way are misapplying their knowledge of whole number addition and subtraction to fraction problems”
MATH-12	Siegler, Carpenter, Fennell, Geary, Lewis, Okamoto, Thompson & Wray 2010 p. 32 (PDF p. 38), Recommendation 3 · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf	The IES guide describes a closely related error: changing a fraction's denominator without making the matching change to the numerator. Its example is fraction addition; applying it to equivalence (our “changed only the bottom” tag) is our inference. “error of changing the denominator of a fraction without making the corresponding change to the numerator”
MATH-13	Siegler, Carpenter, Fennell, Geary, Lewis, Okamoto, Thompson & Wray 2010 p. 31 (PDF p. 37), Recommendation 3 · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf	The IES panel considers it critical to identify students who hold specific fraction misconceptions, to discuss those misconceptions with them, and to make clear why they lead to wrong answers. “it is critical to identify students who are operating with such misconceptions, to discuss the misconceptions with them”
MATH-14	Siegler, Carpenter, Fennell, Geary, Lewis, Okamoto, Thompson & Wray 2010 p. 23 (PDF p. 29), Recommendation 2, step 3 · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf	The IES guide recommends using number lines to teach fraction equivalence: equivalent fractions are the same point (the same magnitude) on the line. “Number lines can be used to illustrate that equivalent fractions describe the same magnitude.”
MATH-15	Siegler, Carpenter, Fennell, Geary, Lewis, Okamoto, Thompson & Wray 2010 p. 20 (PDF p. 26), Recommendation 2, Summary of evidence · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf	Honest limit: the IES panel says its evidence for the number-line recommendation comes mainly from studies of whole numbers (plus one decimals study), not from studies of fractions. “Evidence for this recommendation primarily comes from studies demonstrating the usefulness of number lines for developing number sense with whole numbers.”
MATH-16	Siegler, Carpenter, Fennell, Geary, Lewis, Okamoto, Thompson & Wray 2010 p. 1 (PDF p. 7), Review of Recommendations, Recommendation 3 · opened 2026-10-05 https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/fractions_pg_093010.pdf	The IES fractions guide recommends area models, number lines and other visual representations to help students understand why fraction procedures work (Recommendation 3, rated Moderate Evidence). “Use area models, number lines, and other visual representations to improve students’ understanding of formal computational procedures.”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
MATH-17	Hamilton, Halverson, Jackson, Mandinach, Supovitz & Wayman 2009 p. 10 (PDF p. 16), Recommendation 1 · opened 2026-10-05 https://web.education.wisc.edu/halverson/wp-content/uploads/sites/33/2012/09/What-Works-DDDM-Report-093009.pdf	The IES practice guide on using student data describes a cycle: interpret the data and form hypotheses, change instruction to test them, then collect and interpret new data to evaluate the change. The panel rates its evidence for this recommendation as Low. “Finally, they should restart the cycle by collecting and interpreting new student performance data to evaluate their own instructional changes.”
MATH-18	Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel & Metcalfe 2007 p. 5 (PDF p. 16), Recommendation 1 · opened 2026-10-05 https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/20072004.pdf	The IES practice guide on organizing instruction recommends re-exposing students to key content after a delay of several weeks to several months, because delayed re-exposure often markedly increases how much they remember (Moderate Evidence). “exposed to key course concepts on at least two occasions—separated by a period of several weeks to several months”
FAMILIES, ACCESS AND PRIVACY		
EQ-01	CAST, 2024 Consideration 1.1, implementation examples; accessed 2026-10-05 · opened 2026-10-05 https://udlguidelines.cast.org/representation/perception/customize-display/	CAST's current UDL Guidelines (version 3.0) list adjustable font, text size, spacing, background colour and text colour as ways to let learners customize how information is displayed. “Font, size of text, character and line spacing, character width, background color, and text colors”
EQ-02	CAST, 2024 Consideration 1.2, paragraph on text as visual information; accessed 2026-10-05 · opened 2026-10-05 https://udlguidelines.cast.org/representation/perception/ways-perceive-information/	UDL 3.0 identifies turning text into audio as one of the easiest ways to increase access to text. “The transformation from text into audio is among the most easily accomplished methods for increasing accessibility.”
EQ-03	CAST, 2024 Consideration 4.1, implementation examples; accessed 2026-10-05 · opened 2026-10-05 https://udlguidelines.cast.org/action-expression/interaction/response-navigation-movement/	UDL 3.0 recommends offering several physical ways to respond, including keyboard and switch-style inputs, rather than a single input method. “Offer options for physically interacting with materials by hand, voice, single switch, joystick, keyboard, or adapted keyboard.”
EQ-04	CAST, 2024 Consideration 8.5, implementation examples; accessed 2026-10-05 · opened 2026-10-05 https://udlguidelines.cast.org/engagement/effort-persistence/feedback/	UDL 3.0 recommends feedback that is substantive and informative, not comparative or competitive. “Offer feedback that is substantive and informative rather than comparative or competitive.”
EQ-05	CAST, 2024 Consideration 7.4, implementation examples; accessed 2026-10-05 · opened 2026-10-05 https://udlguidelines.cast.org/engagement/interests-identities/biases-threats-distractions/	UDL 3.0 lists variation in pace, session length and access to breaks as ways to reduce threats and distractions in learning. “Variation in pace of work, length of work sessions, availability of breaks or time-outs, or timing or sequence of activities”
EQ-06	W3C Accessibility Guidelines Working Group, 2024 Introduction, “New Features in WCAG 2.2”; accessed 2026-10-05 · opened 2026-10-05 https://www.w3.org/TR/WCAG22/	WCAG 2.2 has been a W3C Recommendation since 12 December 2024, and the W3C working group recommends adopting 2.2 as the conformance target even where policy names an earlier version. “The Accessibility Guidelines Working Group recommends that sites adopt WCAG 2.2 as their new conformance target, even if formal obligations mention previous versions”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
EQ-07	W3C Accessibility Guidelines Working Group, 2024 Success Criterion 2.5.7 Dragging Movements (AA); accessed 2026-10-05 · opened 2026-10-05 https://www.w3.org/TR/WCAG22/#dragging-movements	WCAG 2.2 (Level AA) requires that anything done by dragging can also be done with a single pointer without dragging. “All functionality that uses a dragging movement for operation can be achieved by a single pointer without dragging”
EQ-08	W3C Accessibility Guidelines Working Group, 2024 Success Criterion 3.3.8 Accessible Authentication (Minimum) (AA); accessed 2026-10-05 · opened 2026-10-05 https://www.w3.org/TR/WCAG22/#accessible-authentication-minimum	WCAG 2.2 (Level AA) says sign-in must not depend on a cognitive function test, such as remembering a password, unless an alternative or assisting mechanism is provided; the criterion lists further exceptions (object recognition, personal content). “A cognitive function test (such as remembering a password or solving a puzzle) is not required for any step in an authentication process”
EQ-09	Federal, Provincial and Territorial Privacy Commissioners and Ombuds, 2025 Calls to school boards, post-secondary institutions and administrators; accessed 2026-10-05 · opened 2026-10-05 https://www.priv.gc.ca/en/about-the-opc/what-we-do/provincial-and-territorial-collaboration/joint-resolutions-with-provinces-and-territories/res_20251008_edtech/	In October 2025, Canada’s federal, provincial and territorial privacy regulators jointly called on school boards to use privacy-protective settings by default in EdTech and to support pseudonyms, generic logins or anonymous profiles. “ensure that products and services are secure, use privacy protective settings by default, and support the use of pseudonyms and generic logins or anonymous profiles”
EQ-10	Federal, Provincial and Territorial Privacy Commissioners and Ombuds, 2025 Calls to EdTech vendors; accessed 2026-10-05 · opened 2026-10-05 https://www.priv.gc.ca/en/about-the-opc/what-we-do/provincial-and-territorial-collaboration/joint-resolutions-with-provinces-and-territories/res_20251008_edtech/	The same 2025 joint resolution asks EdTech vendors to collect, use, disclose or keep children’s personal information only as necessary and proportionate for delivering educational services. “Only collect, use, disclose or retain the personal information of children and youth as is necessary and proportionate for the delivery of educational services.”
EQ-11	Federal, Provincial and Territorial Privacy Commissioners and Ombuds, 2025 Context, “Profiling and inferences”; accessed 2026-10-05 · opened 2026-10-05 https://www.priv.gc.ca/en/about-the-opc/what-we-do/provincial-and-territorial-collaboration/joint-resolutions-with-provinces-and-territories/res_20251008_edtech/	The 2025 joint resolution names profiling and inference as a key EdTech risk, including predictions about performance that could shape a child’s educational path in advance. “make predictions about performance which could impact a child’s classroom experience and pre-emptively shape their educational path”
EQ-12	Federal, Provincial and Territorial Privacy Commissioners and Ombuds, 2025 Calls to EdTech vendors; accessed 2026-10-05 · opened 2026-10-05 https://www.priv.gc.ca/en/about-the-opc/what-we-do/provincial-and-territorial-collaboration/joint-resolutions-with-provinces-and-territories/res_20251008_edtech/	The 2025 joint resolution asks EdTech vendors to use clear, accessible language so that children and youth can understand what is done with their personal information. “Use clear and accessible language for all users, especially children and youth, so that they can understand what is being done with their personal information.”
EQ-13	Information and Privacy Commissioner of Ontario, 2024 Single page, “Protection” pledge item 5; accessed 2026-10-05 · opened 2026-10-05 https://www.ipc.on.ca/sites/default/files/documents/2024-07-04-Digital%20Privacy%20Charter-8pt5x14.pdf	The Information and Privacy Commissioner of Ontario’s Digital Privacy Charter for Ontario Schools includes a pledge to make privacy-protective settings the default in digital education tools. “Strengthening privacy protections in digital education tools and services, including by making privacy-protective settings the default when using them”

CLAIM	SOURCE, LOCATION, LINK	WHAT WE CITE IT FOR, AND THE SUPPORTING QUOTATION
EQ-14	U.S. Department of Education, Privacy Technical Assistance Center, 2014 p. 9, "Data Access Provisions"; accessed 2026-10-05 · opened 2026-10-05 https://studentprivacy.ed.gov/sites/default/files/resource_document/file/Student%20Privacy%20and%20Online%20Educational%20Services%20%28February%202014%29_0.pdf	US Department of Education guidance recommends that parents request records held by a provider through the school or district, which obtains them from the provider, instead of the provider authenticating parents itself. "the Department recommends that the school or district serve as the intermediary for these requests"
EQ-15	Federal Trade Commission (page notes the Rule was amended April 22, 2025) Section A.3, "What is Personal Information?"; accessed 2026-10-05 · opened 2026-10-05 https://www.ftc.gov/business-guidance/resources/complying-coppa-frequently-asked-questions	Under the US COPPA Rule as described by the FTC (which concerns children under 13), personal information includes a persistent identifier that can recognize a user over time and across different websites or online services. The guidance does not say whether a classroom alias meets this definition; 1080 treats aliases as personal information as a precaution. "A persistent identifier that can be used to recognize a user over time and across different websites or online services"
EQ-16	Kraft & Rogers, 2015 Abstract; accessed 2026-10-05 · opened 2026-10-05 https://ideas.repec.org/a/eee/ecoedu/v47y2015icp49-63.html	In a randomized field experiment, weekly one-sentence individualized teacher messages to parents reduced the share of students failing to earn credit, mainly by preventing drop-outs, and messages about what students could improve had the largest effects (as described in the published abstract). "messages emphasizing what students could improve, versus what students were doing well, producing the largest effects"
EQ-17	Kraft, 2016 p. 15 (PDF p. 1); accessed 2026-10-05 · opened 2026-10-05 https://www.bostonfed.org/-/media/Documents/cb/The-Underutilized-Potential-of-Teacher-Parent-Communication.pdf	Kraft reports, from US national survey data, that school email to families at or below the poverty line stayed flat after 2007 while the income gap in email contact widened. "Email communication with families living at or below the poverty line has remained flat since 2007."
EQ-18	York & Loeb, 2014 (rev. 2018) PDF p. 13, intervention description; abstract states "increased parental involvement at home and school by 0.15 to 0.29 standard deviations"; accessed 2026-10-05 · opened 2026-10-05 https://www.nber.org/system/files/working_papers/w20659/w20659.pdf	The READY4K! text-message program for parents of preschoolers sent short, simple, highly specific activities built on existing family routines; its randomized evaluation reports gains in parental involvement and early literacy (see abstract at nber.org/papers/w20659). "These texts include short, simple, and highly specific activities for parents to do with their children that build on existing family routines."



About this edition

Research Edition, version 1.0, 5 October 2026. Written by Brandon Robert, Ontario Certified Teacher (registered with the Ontario College of Teachers under his legal name), Milestone Teachers. Questions and corrections: admin@milestoneteachers.ca. The web edition, with the same text and figures, is at milestoneteachers.com/1080-research/.

Figures are drawn from the public demo's own content and verified keys. A build check recomputes every item value in the figures and in the worked-example table (each key, each tagged answer and the path counts), and it was shown to fail on planted errors before this edition was made.

Type and images

Set in Instrument Serif, Instrument Sans and IBM Plex Mono, all under the SIL Open Font License. Cover and plate images are rendered still lifes from the Optical Bench series made for Milestone Teachers. They show glass, stone and light only: no real people, classrooms or learner data.

Learners named in this paper (Heron, Cedar, Delta) are sample.