Diffit Quality Rubric

A scoring rubric for evaluating AI-generated instructional materials, derived from the Diffit Quality Constitution and the WestEd four-category framework for instructional materials quality (Bugler et al., 2017).

How to use this rubric

For each criterion, score the artifact:

- **Meets** — clearly satisfies the commitment on this artifact, with evidence
- **Partial** — partly satisfies; minor gaps a teacher could correct in seconds
- **Fails** — does not satisfy; a teacher would have to substantially rework

Every score must cite specific evidence — a quote, a question number, a page reference, a screenshot moment. Unsourced scores don't belong in this comparison.

N/A is honest. If a criterion doesn't apply to the prompt (e.g., source fidelity when no source was provided), mark N/A rather than forcing a score.

Diffit will not score perfectly. The Constitution is forward-looking; some commitments are aspiration. We score Diffit honestly and let the gap to competitors speak.

---

1. Accuracy and Craft

> The facts are right, the layout is clean, and a teacher will not be embarrassed in front of the class.

1.1 Factual correctness

Are facts, numbers, dates, names, units, and scientific claims accurate at a level a student at the named grade can verify? Does math compute?

Score	Anchor
Meets	Spot-check finds no factual errors in passage, vocab definitions, questions, or answer key
Partial	One minor, non-load-bearing inaccuracy (e.g., a date is the year before the actual year)
Fails	A factual claim a student or family could catch — wrong process, wrong number, wrong scientific mechanism

1.2 Source fidelity

When a source is provided (URL, PDF, text), does the material work *from* it without invention or substitution? Are quotations and data accurate to the source?

Score	Anchor
Meets	Every substantive claim traces to the source; no facts smuggled in from elsewhere
Partial	Mostly source-grounded; one or two details added or implied beyond the source
Fails	Invents details not in the source, fabricates quotes, substitutes a different source, or rewrites the problem
N/A	No source provided in the prompt

1.3 Mechanical correctness

Spelling, grammar, punctuation, capitalization. The boring stuff.

Score	Anchor
Meets	Clean enough to hand a student as a final draft
Partial	1–2 minor errors a teacher would not flag in a colleague's draft
Fails	Visible errors a teacher must correct before assigning

1.4 Layout discipline (page density, graphics, ink)

Does the page earn its length? Density, whitespace, font sizing scaled to grade. Graphics only when they do pedagogical work, not for decoration.

Score	Anchor
Meets	Density matches grade band; no padding; every graphic is informative
Partial	Usable but visually inefficient (e.g., 3 pages of what fits on 1; one decorative graphic)
Fails	Wall of text, padded with filler, or decorative graphics that distract

1.5 Answer-key consistency

Does the answer key actually answer the questions in *this* artifact, in the form the questions imply?

Score	Anchor
Meets	Every answer matches a question in this artifact; form (fraction/decimal, full sentence, etc.) matches
Partial	Most match; one or two minor drifts in form or numbering
Fails	Missing answers, wrong questions answered, or no key produced when one was expected
N/A	No answer key requested or implied

---

2. Standards Alignment and Depth of Knowledge

> The material hits the standard the teacher named, at the depth that standard requires — not a watered-down version of it.

2.1 Standards alignment

Is a specific standard cited (NGSS, CCSS, state-specific), and does the content actually meet it (not just "cover the topic")?

Score	Anchor
Meets	Explicit standard cited, and content addresses the verbs in the standard (e.g., NGSS 5-ESS2-1 asks for a model — the artifact has students build one)
Partial	Topic-correct, but no standard cited, or weakly aligned
Fails	No standard cited and content drifts off-topic, or claims a standard but does not meet its cognitive demand

2.2 Cognitive depth (DOK spread)

Across the questions in the packet, is there a real spread of cognitive demand, or all recall?

Score	Anchor
Meets	At least one question at DOK 3+ (analyze, evaluate, construct argument from evidence) alongside foundational recall
Partial	Mostly recall with one stretch question
Fails	All recall / one cognitive level only

2.3 Genuine variety

Do different question *formats* (MC, short answer, open-ended, matching) ask for meaningfully different *thinking*, or just the same recall in different visual layouts?

Score	Anchor
Meets	The thinking varies across questions, not just the format
Partial	Format varies; thinking varies somewhat
Fails	Same shallow recall asked three different ways

2.4 Question quality

Are multiple-choice distractors plausible, related to the content, and educative (revealing common misconceptions)? Does the position of the correct answer vary?

Score	Anchor
Meets	Distractors are content-relevant and educative; correct-answer positions vary across the set
Partial	One of the two: educative distractors but predictable positions, or varied positions but throwaway distractors
Fails	Throwaway distractors ("none of the above," nonsense options), or "always B" patterns

2.5 Honesty about what the standard requires

When a standard asks for extended writing, modeling, or argument from evidence, does the artifact actually demand that work — or is it satisfied with a fill-in-the-blank?

Score	Anchor
Meets	Where the standard demands depth, the artifact provides it (essay prompts, modeling tasks, evidence-based argument prompts)
Partial	Some depth, but the hardest cognitive moves are routed around
Fails	A standard requiring extended writing or modeling is satisfied with recall-level tasks
N/A	No standard cited

---

3. Ease of Use and Completeness

> A teacher can walk into the classroom and use this. No missing pieces, no internal inconsistencies, no "the teacher should now explain..." placeholders.

3.1 Multi-activity coherence

When the packet contains a passage + quiz + vocab + key, do they reference the same content — same details, same vocabulary, same emphasis?

Score	Anchor
Meets	The packet reads as a single resource; quiz asks about this passage; vocab drawn from this passage
Partial	Mostly coherent; minor drift (e.g., vocab not all drawn from passage)
Fails	Questions ask about content not in the passage; vocab unrelated; pieces feel independently generated

3.2 Complete and ready to teach

Can a teacher use the artifact as-is without hunting for missing pieces or filling in placeholders?

Score	Anchor
Meets	No placeholders, no "[insert passage here]," no "the teacher should now explain..."
Partial	Minor gaps a teacher fixes in under a minute
Fails	Stage directions, placeholders, or instructions to "find a source and fill it in"

3.3 Considered defaults

Does a single, simple prompt produce a usable output? Or does the teacher need to know the right follow-up prompts to get there?

Score	Anchor
Meets	One prompt, usable output. Early-career teacher could ship it
Partial	Usable with 1–2 obvious follow-ups
Fails	Requires significant prompt engineering / iteration to get to "usable"

3.4 Honest layout for student work

Does the space for student writing match what they're being asked to produce? (Essay → real lines; one-word answer → one line; show-your-work math → room for steps.)

Score	Anchor
Meets	Writing space appropriate to each task
Partial	One task miscalibrated (e.g., 3 lines for an essay)
Fails	Multiple tasks miscalibrated, or no writing space provided where it's needed

3.5 Classroom-workflow format fidelity

Does the output land cleanly in the formats teachers actually use — print-ready PDF, Google Docs, Google Slides, Forms?

Score	Anchor
Meets	Exports / prints cleanly to a teacher-ready file in at least two formats
Partial	Usable in one format; breaks in others
Fails	Copy-paste artifact; formatting collapses; no export path

---

4. Engagement and Meeting Student Needs

> The materials meet students where they are — reading level, language, prior knowledge, cultural context — and engage them rather than droning at them.

4.1 Right reading level

Sentence length, clause complexity, vocabulary, and conceptual density match the named grade.

Score	Anchor
Meets	Reads like the named grade — short clear sentences for younger, compound sentences and academic vocabulary (introduced with context) for older
Partial	Slightly above or below the named grade
Fails	Clearly miscalibrated — either dumbed-down (deleting content) or unreadable for the grade

4.2 Engaging voice

Does the passage hold a student's attention, or read as generic "AI-written" textbook drone?

Score	Anchor
Meets	Concrete examples, varied sentence structure, a voice a student would not skim past
Partial	Serviceable but flat
Fails	Generic, repetitive, or filler ("In this passage, we will learn about...")

4.3 Differentiation in the workflow

Can the same lesson be produced at multiple reading levels / languages / scaffolds from the same brief, without rebuilding?

Score	Anchor
Meets	Multiple differentiated versions delivered or one click away (e.g., 3 reading levels, ELL scaffold, translation)
Partial	Differentiation possible but requires separate, manually-crafted prompts
Fails	One version only; no path to differentiation in the workflow

4.4 Pedagogically-useful visuals

For topics where a diagram carries conceptual load (water cycle, photosynthesis, geometry, data interpretation), is the visual present, accurate, and grade-appropriate?

Score	Anchor
Meets	Diagram present, accurately labeled, scientifically correct, clarifies the concept
Partial	Present but generic, or partly inaccurate
Fails	No diagram for a topic that needs one; or a diagram that misrepresents the science / has hallucinated text artifacts

4.5 Asset-based framing and safety

Do examples, names, and contexts draw from broad cultural backgrounds rather than assuming a default student? Is content age-appropriate?

Score	Anchor
Meets	Examples draw from broad contexts; no default-student framing; nothing age-inappropriate
Partial	Neutral but bland; minor default-student assumptions
Fails	Excluding / stereotyping defaults, or age-inappropriate content

---

Scoring summary template

A side-by-side comparison should end with a heatmap like this:

Criterion	Diffit	Competitor
1.1 Factual correctness	✓	✓
1.2 Source fidelity	✓	⚠
…	…	…

Followed by:

- **Net assessment** — one paragraph
- **3–5 "look here" moments** — the specific evidence callouts an admin should look at

- **What we don't claim** — criteria where Diffit underperforms or where the difference is small. Honesty is the differentiator.

Citation

This rubric operationalizes the Diffit Quality Constitution, itself drawing on:

Bugler, D., Marple, S., Burr, E., Chen-Gaddini, M., & Finkelstein, N. (2017). *How Teachers Judge the Quality of Instructional Materials: Selecting Instructional Materials, Brief 1 — Quality.* WestEd. https://www.wested.org/resource/selecting-instructional-materials-brief-1-quality/

Diffit vs. Brisk

A side-by-side comparison of two AI tools *built for educators*, each generating the same instructional packets across grade bands, scored against the Diffit Quality Rubric — which is derived from the Diffit Quality Constitution and WestEd's research on how teachers judge instructional-materials quality.

Audience: District and school leaders who already have Brisk and want to understand what Diffit adds.

Approach: Three identical prompts, each run once in each tool, scored on observable artifact qualities. Every Brisk artifact was produced with the tool's **Quick Create** flow (web app), entering the topic, grade, standard code (selected by hand), and, where applicable, the source URL, then generating once with no follow-up. Brisk generates into a Google Doc.

Why this is a closer contest than a general-purpose chatbot. Brisk is purpose-built for teachers, and it shows. Its math is accurate, its output lands in a clean Google Doc, it accepts a source URL, and — this is the real strength — **it writes genuinely good, standards-aware questions**. We credit all of that explicitly. The gap to Diffit opens one layer deeper: Brisk writes good questions but does not build the lesson *around* them. On the source-grounded prompts it links or shrinks the source instead of building a passage from it; on math it omits the visual models the standard names; and its answer key appears on one worksheet but not the others.

How this maps to the summary page: the four criteria shown there — *Classroom-ready & standards-aligned*, *Complete & ready to teach*, *Content integrity*, and *Differentiation that holds up* — are presentation themes drawn from the 20-criterion rubric. This document is the full per-criterion scoring behind them.

TL;DR

When a teacher asks each tool for the same packet, **Brisk reliably writes thoughtful, standards-aware questions in a clean Google Doc — but it does not build the lesson around them. Diffit treats the standard and the source as inputs that shape a complete packet: a passage built from the source, activities that meet the standard's cognitive demand, visual models where the standard names them, and a consistent answer key.**

	Diffit	Brisk
Math packet (fractions, 5.NF.A.2)	Visual fraction models + unlike-denominator word problems with a required estimate each + an error-analysis activity + a full worked answer key	Accurate and clean, with worked examples and an estimation tips box — but text-only (no visual models the standard names), "show your work" with a single inline blank, and no answer key
ELA packet (Harriet Tubman, NPS source, RI.5.3)	Reading passage built from the source + 3 text-based activities + 2 images + answer key that quotes the source	Genuinely good RI.5.3 questions and a teacher key — but **no passage** (Part 1 is a hyperlink), questions ask for "an example from the text" that isn't there, and the key is generic (no NPS specifics)
HS packet (MLK Letter from Birmingham Jail, RH.9-10.2)	Built from the whole letter + 8 activity types + a source-grounded key quoting the letter	The entire "excerpt" is a single 40-word quote; the four vocabulary words are the nouns in it; no answer key
Considered defaults (one-prompt-to-usable)	✓ across all three prompts	✓ — Quick Create returns a usable worksheet in one shot every time
Differentiation (2nd-grade relevel)	Whole packet releveled — passage included, names/dates/story kept — in one step	Questions and answers releveled cleanly (credit), but the reading stays the adult NPS link
Pattern overall	Wins on source fidelity, standards depth, answer-key consistency, visual models, and room to write	Clears the hygiene bar (accurate, clean, Google Docs, good questions) but does not build the lesson around its questions

The single most important finding: a worksheet is more than its questions. Brisk writes good, standards-aware questions — but on the source-grounded prompts it links or shrinks the source instead of building from it, on math it omits the visual models the standard names, and its answer key shows up on one worksheet and not the others. Diffit builds the whole packet from the standard and the source.

We score Diffit honestly against its own (aspirational) Constitution and call out where Diffit falls short — a stray rendering artifact on the math sheet, a single-item oversimplification in the 2nd-grade relevel — so the wins it does post land with more weight. And we credit Brisk everywhere it earns it: accurate math, clean Google-Docs output, well-designed questions, and a genuinely good worksheet relevel.

What we tested

Three prompts spanning two grade bands, all CCSS-aligned.

Prompt 1 — Math

> **Topic:** Adding and subtracting fractions with unlike denominators **Grade:** 5 **Standard:** CCSS.MATH.CONTENT.5.NF.A.2 — *"Solve word problems involving addition and subtraction of fractions referring to the same whole, including cases of unlike denominators, e.g., by using visual fraction models or equations to represent the problem."* **Source:** none (the standard is the brief)

Prompt 2 — ELA

> **Topic:** Harriet Tubman **Grade:** 5 **Standard:** CCSS.ELA-LITERACY.RI.5.3 — *"Explain the relationships or interactions between two or more individuals, events, ideas, or concepts in a historical, scientific, or technical text based on specific information in the text."* **Source:** NPS — Harriet Tubman

Prompt 3 — HS ELA

> **Topic:** Martin Luther King Jr.'s "Letter from Birmingham Jail" **Grade:** 9–10 **Standard:** CCSS.ELA-LITERACY.RH.9-10.2 — *"Determine the central ideas or information of a primary or secondary source; provide an accurate summary of how key events or ideas develop over the course of the text."*
> **Source:** Stanford King Institute encyclopedia entry on the letter

---

The primary artifacts

Artifact	File	Pages
Diffit Math — fractions	diffit-math-fractions.pdf	15
Brisk Math — fractions (Quick Create → Worksheet)	brisk-math-fractions.pdf	4
Diffit ELA — Harriet Tubman	diffit-ela-tubman-v2.pdf	9
Brisk ELA — Harriet Tubman (Quick Create → Worksheet)	brisk-tubman-worksheet.pdf	3
Diffit HS — MLK Letter	diffit-ela-mlk.pdf	17
Brisk HS — MLK Letter (Quick Create → Worksheet)	brisk-mlk-worksheet.pdf	3
Diffit ELA — Tubman, releveled to Grade 2	diffit-ela-tubman-2nd-grade.pdf	10
Brisk ELA — Tubman, releveled to Grade 2 (Change Level)	brisk-2nd-grade.pdf	3

Note: Brisk's first response was usable in every case — no follow-up prompting was required to get a student-facing worksheet. That is a genuine §3.3 (considered defaults) credit, and it holds across all three prompts. (A Brisk “Resource Pack” of the math topic was also generated as a cross-check; the dedicated Worksheet output above is the fairer student-facing comparison and is the one scored.)

Math scoring

Diffit Math vs Brisk Math (fractions), against the rubric:

#	Criterion	Diffit	Brisk	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	Every computation checks out on both sides. Diffit's worked key is correct throughout; Brisk's two worked examples and all practice and word-problem answers compute correctly (1/4 + 2/3 = 11/12, 5/6 − 1/4 = 7/12, and so on).
1.2	Source fidelity	N/A	N/A	No source provided — the standard is the brief.
1.3	Mechanical correctness	⚠ Partial	✓ Meets	Brisk is mechanically clean. Diffit's error-analysis sheet carries a stray rendering artifact — "½ + $14" where "½ + ¼" was meant — a minor, non-load-bearing slip, scored honestly. This is the one criterion where Brisk edges Diffit.
1.4	Layout discipline	✓ Meets	✓ Meets	Both are clean and appropriately dense; neither is padded. Brisk exports as a tidy Google Doc.
1.5	Answer-key consistency	✓ Meets	× Fails	Diffit includes a full worked answer key. Brisk's worksheet ships with no answer key and offers no toggle to add one — the teacher solves all twelve problems.
2.1	Standards alignment	✓ Meets	✓ Meets	Both cite 5.NF.A.2. Diffit's tasks use visual models and unlike-denominator word problems; Brisk's word problems and equation practice address the standard's verbs. Credit Brisk here.
2.2	Cognitive depth (DOK)	✓ Meets	⚠ Partial	Diffit includes a dedicated error-analysis activity (DOK 3) and a required estimate on each word problem. Brisk is mostly procedural — compute the sum or difference — with word-problem application but no reasoning-about-reasonableness task.
2.3	Genuine variety	✓ Meets	⚠ Partial	Diffit's models / solve / estimate / critique each demand different thinking. Brisk varies the wrapper (worked example, bare practice, word problem) but the thinking is the same compute-the-fraction step.
2.4	Question quality	N/A	N/A	Neither worksheet uses multiple choice.
2.5	Honesty about what the standard requires	✓ Meets	⚠ Partial	5.NF.A.2 names visual fraction models and the use of benchmark fractions to estimate and assess reasonableness. Diffit requires an estimate on every problem plus a misconception analysis; Brisk mentions estimation in a tips sidebar (advisory, not a task) and provides no visual models — the standard's reasoning move is mentioned, not demanded.
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both hang together around a consistent fractions context.
3.2	Complete and ready to teach	✓ Meets	⚠ Partial	Diffit is complete with a key. Brisk's worksheet is student-ready as a handout, but the teacher must produce the answer key.

#	Criterion	Diffit	Brisk	Evidence
3.3	Considered defaults	✓ Meets	⚠ Partial	Both yield a usable worksheet from one prompt. Brisk's default omitted the key — an obvious follow-up to reach fully usable.
3.4	Honest layout for student work	✓ Meets	× Fails	Brisk instructs students to “show your work by finding the common denominator and simplifying,” then provides a single inline blank (“= ____”); the word problems have no work space at all. Diffit sizes ruled space to each task.
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both export and print cleanly — Brisk lands natively in Google Docs.
4.1	Right reading level	✓ Meets	✓ Meets	Word problems read at 5th-grade level in both.
4.2	Engaging voice	✓ Meets	⚠ Partial	Diffit uses “Fraction Freddy” and named scenarios. Brisk's prose is functional but flat.
4.3	Differentiation in the workflow	N/A	N/A	Differentiation tested in its own section below.
4.4	Pedagogically-useful visuals	✓ Meets	× Fails	5.NF.A.2 names visual fraction models in the standard text. Diffit includes bar, circle, and fraction-bar models with shading tasks. Brisk's worksheet is text-only — a teacher using it to fulfill 5.NF.A.2 has not delivered the representation the standard names.
4.5	Asset-based framing	✓ Meets	✓ Meets	Both name a diverse set of students (Diffit: Maya, Leo, Sarah, Chloe; Brisk: Maria, Liam).

Math scorecard: Diffit 16 Meets, 1 Partial, 0 Fails (out of 17 applicable; 1.2, 2.4, 4.3 N/A). Brisk 8 Meets, 6 Partials, 3 Fails (out of 17 applicable). The load-bearing Fails are §4.4 (no visual models, which the standard names), §3.4 (no room to show the work it asks for), and §1.5 (no answer key). Diffit's lone non-Meet is the stray “$” rendering artifact on the error-analysis sheet (§1.3).

ELA scoring

Diffit ELA vs Brisk ELA (Harriet Tubman), against the rubric:

#	Criterion	Diffit	Brisk	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	Both are factually accurate. Brisk's answer-key facts (escaped slavery, Underground Railroad conductor, Civil War spy and nurse) are correct, if general.
1.2	Source fidelity	✓ Meets	× Fails	Diffit works from the NPS page in specific detail — Araminta Ross, Edward Brodess, the Bucktown Village Store head injury, the Parson's Creek mariners, the Combahee River Raid (June 1, 1863; 750 freed) — and its answer key quotes those specifics. Brisk was given the same NPS URL but produced no passage, and its answer key is generic: none of the source's specifics appear, evidence the URL was linked rather than read.
1.3	Mechanical correctness	✓ Meets	⚠ Partial	Diffit is clean. Brisk's Part 1 contains a broken Markdown link — the literal text "[Harriet Tubman - NPS](" with a stray ")" below the URL.
1.4	Layout discipline	✓ Meets	⚠ Partial	Diffit's packet is well-structured. Brisk's "Part 1: Reading Comprehension" is a near-empty section (a single link) — a structural hollow.
1.5	Answer-key consistency	✓ Meets	✓ Meets	Both include an answer key whose answers match the questions. Credit Brisk — this worksheet's "Examples of Answers" covers all five questions. (The generic-content issue is scored under 1.2, not here.)
2.1	Standards alignment	✓ Meets	⚠ Partial	Both name RI.5.3, and Brisk's questions genuinely target the standard's verbs (relationships, interactions, connections). But RI.5.3 requires explanation "based on specific information in the text," and Brisk supplies no text — so the close reading the standard names cannot be grounded in the artifact.
2.2	Cognitive depth (DOK)	✓ Meets	✓ Meets	Brisk's questions reach analysis and connection (DOK 2–3), with a short-writing paragraph and a comparison challenge. Credit the question design.
2.3	Genuine variety	✓ Meets	✓ Meets	Both field genuinely different thinking — analysis questions, extended writing, comparison.
2.4	Question quality	N/A	N/A	Neither worksheet uses multiple choice.
2.5	Honesty about what the standard requires	✓ Meets	⚠ Partial	Diffit's interaction and cause/effect tables demand the text-based analysis the standard names. Brisk demands analysis in the abstract but routes the "from the text" move to an external link.
3.1	Multi-activity coherence	✓ Meets	× Fails	Diffit's passage, activities, and key reference the same content. Brisk's Question 4 asks for "an example from the text" — but no text is in the worksheet; the pieces do not cohere.

#	Criterion	Diffit	Brisk	Evidence
3.2	Complete and ready to teach	✓ Meets	✗ Fails	Diffit is a complete, self-contained packet. Brisk's is not usable as handed out: “Part 1: Reading Comprehension” is a hyperlink to the adult NPS page, so a student needs the internet open and must read an adult website.
3.3	Considered defaults	✓ Meets	✗ Fails	Diffit's single prompt yields a complete packet. Brisk's single prompt yields a reading worksheet with no reading — a teacher must source, level, and paste a passage to make it usable, which is not a quick fix.
3.4	Honest layout for student work	✓ Meets	⚠ Partial	Diffit sizes ruled space to each task. Brisk gives “Answer:” lines on the questions but no space for the Part 3 paragraph it assigns.
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both export and print cleanly — Brisk into Google Docs.
4.1	Right reading level	✓ Meets	✓ Meets	Brisk's questions read at grade level (there is no passage to assess).
4.2	Engaging voice	✓ Meets	⚠ Partial	Diffit's passage uses concrete narrative detail. Brisk has no passage — the worksheet is questions only, so there is no reading experience to engage a student.
4.3	Differentiation in the workflow	N/A	N/A	Differentiation tested in its own section below.
4.4	Pedagogically-useful visuals	N/A	N/A	A biography reading does not require a conceptual diagram; marked N/A for both. (Diffit's packet does include two contextual images.)
4.5	Asset-based framing	✓ Meets	✓ Meets	Both handle the subject respectfully and accurately.

ELA scorecard: Diffit 17 Meets, 0 Partials, 0 Fails (out of 17 applicable; 2.4, 4.3, 4.4 N/A). Brisk 7 Meets, 6 Partials, 4 Fails (out of 17 applicable). The four Fails (§1.2, §3.1, §3.2, §3.3) all trace to a single root cause: the worksheet has no reading passage. Credit where due — Brisk's RI.5.3 questions are well-designed (§2.2, §2.3 Meet) and this worksheet does include a key (§1.5 Meets).

HS scoring: Letter from Birmingham Jail

Diffit HS vs Brisk HS (MLK), against the rubric:

#	Criterion	Diffit	Brisk	Evidence
1.1	Factual correctness	✓ Meets	✓ Meets	Both accurate. Brisk's single quotation is transcribed correctly, and the historical context it supplies (jailed for protesting segregation, responding to local clergy) is right.
1.2	Source fidelity	✓ Meets	⚠ Partial	Diffit builds from the actual letter — the four-step campaign, the clergy's charges, King's own words — and its answer key quotes the source. Brisk's one quote is real and the Stanford source is cited, so nothing is fabricated or substituted; but the worksheet works from roughly 40 words of a ~7,000-word source, so it barely engages the source at all.
1.3	Mechanical correctness	✓ Meets	⚠ Partial	Diffit is clean. Brisk's source line is a broken Markdown link — the literal "[The Martin Luther King, Jr. Research and Education Institute, Stanford University](" with a stray ")".
1.4	Layout discipline	✓ Meets	⚠ Partial	Diffit's multi-worksheet packet is well-structured. Brisk's three pages are built on a single 40-word quote, leaving the sheet thin with large stretches of whitespace.
1.5	Answer-key consistency	✓ Meets	✕ Fails	Diffit includes a detailed, source-grounded answer key — every multiple-choice answer carries a "Source Reference" quoting the letter. Brisk's worksheet has no answer key.
2.1	Standards alignment	✓ Meets	⚠ Partial	Both name RH.9-10.2, which asks students to summarize how key ideas "develop over the course of the text." Diffit traces that development (criticism → counter-argument, the rhetoric of extremism, the myth of inevitability). Brisk's questions target central idea and summary — but on a single sentence, development cannot be traced.
2.2	Cognitive depth (DOK)	✓ Meets	⚠ Partial	Brisk's prompts (explain in your own words, a real-world application, why King responded to critics) are reasonable, but all are bounded by the 40-word quotation.
2.3	Genuine variety	✓ Meets	⚠ Partial	Brisk varies format (questions, vocabulary matching, paragraph, extended thinking) but every item orbits the same quotation. Diffit fields eight distinct activity types across the letter.
2.4	Question quality	N/A	N/A	Brisk's worksheet uses no multiple choice. (Diffit includes a multiple-choice comprehension quiz with varied answer positions; not scored here, to keep a shared denominator.)
2.5	Honesty about what the standard requires	✓ Meets	✕ Fails	Analyzing how an argument develops requires the developing argument — the letter. A single quotation cannot show development, so the standard's central demand is routed around by shrinking the text.

#	Criterion	Diffit	Brisk	Evidence
3.1	Multi-activity coherence	✓ Meets	✓ Meets	Both are internally coherent — Brisk's questions and vocabulary all reference its one quotation.
3.2	Complete and ready to teach	✓ Meets	⚠ Partial	Diffit is a complete packet with a key. Brisk's worksheet is usable as a handout but thin and unkeyed.
3.3	Considered defaults	✓ Meets	⚠ Partial	Both produce a usable artifact from one prompt; a teacher would want the full letter rather than a single quote — an obvious follow-up.
3.4	Honest layout for student work	✓ Meets	⚠ Partial	Diffit sizes space to each task. Brisk's open prompts (“write your answer in 2–3 sentences”) are not given dedicated writing space.
3.5	Classroom-workflow format fidelity	✓ Meets	✓ Meets	Both export and print cleanly — Brisk into Google Docs.
4.1	Right reading level	✓ Meets	✓ Meets	Both pitched at a high-school academic register.
4.2	Engaging voice	✓ Meets	⚠ Partial	Diffit's activities draw on the letter's rhetoric. Brisk's worksheet is questions about a quotation — no reading experience.
4.3	Differentiation in the workflow	N/A	N/A	Differentiation tested in its own section below (on the Tubman packet).
4.4	Pedagogically-useful visuals	N/A	N/A	A text-only primary source does not require a conceptual diagram; marked N/A for both. (Diffit's packet does include flow-map diagrams of the four-step campaign and a cause/effect map.)
4.5	Asset-based framing	✓ Meets	✓ Meets	Both handle civil-rights content with appropriate seriousness.

HS scorecard: Diffit 17 Meets, 0 Partials, 0 Fails (out of 17 applicable; 2.4, 4.3, 4.4 N/A). Brisk 5 Meets, 10 Partials, 2 Fails (out of 17 applicable). The two Fails are §1.5 (no answer key) and §2.5 (the standard's demand routed around); the ten Partials reflect a worksheet built on a single 40-word quotation rather than the letter. Nothing here is fabricated — the quote is real and cited — which is why this is reduction, not invention.

Differentiation in practice

Starting from the same 5th-grade Tubman worksheet, Brisk's **Change Level** tool was asked to relevel it to Grade 2; Diffit was asked for the same. Both are one-action workflows.

Credit Brisk

The relevel did a clean job on the worksheet: the learning goal, the questions, and the sample answers all dropped to a natural Grade-2 voice, and the RI.5.3 relationship framing held. On the worksheet itself, Brisk's relevel is genuinely good.

The gap is the reading

Brisk's worksheet never contained a passage — only the adult NPS link — so the relevel had nothing to lower, and the Grade-2 worksheet still points a second grader at an adult website. A teacher can close that gap by hand (paste the NPS text in, relevel it separately, drop it into the doc), but it is a second pass.

Diffit relevels the whole packet at once. The passage is rewritten for Grade 2 — keeping Araminta Ross, Edward Brodess, the marshland skills, and the Combahee River Raid in simpler sentences — alongside the releveled activities, with images and a Grade-2 answer key. The Constitution's commitment is explicit: *"We do not 'dumb down' by deleting content. We find the simpler way to express the harder idea."* Diffit's relevel keeps the content; Brisk's leaves the content that matters most — the reading — un-leveled.

Honest findings on the Diffit side

Diffit's 2nd-grade relevel is not flawless: one Cause/Effect prompt flattens the Combahee River Raid to *"helps soldiers in a war. They go on a boat trip"* — the 750 freed are still in the answer key, but a 2nd grader reading only that prompt gets a thin picture. A single-item oversimplification, not a whole-relevel failure. And on the math packet, the stray "$" rendering artifact noted under §1.3 is a real craft slip.

What this means for §4.3: Diffit **Meets**, Brisk **Partial**. The differentiation workflow exists and runs cheaply in both, and Brisk's question relevel is genuinely good — but a relevel that leaves the reading at an adult level is incomplete for a reading worksheet.

The five "look here" moments

If a district admin reads nothing else, these are the moments to look at.

1. A "Reading Comprehension" worksheet with no reading

Brisk was given the NPS Harriet Tubman URL. Its “Part 1: Reading Comprehension” is a single hyperlink to that adult page — there is no passage in the worksheet. Question 4 then asks students for “an example from the text,” and the answer key is generic (no Edward Brodess, no Combahee River Raid, no Bucktown store). The source was linked, not read. Diffit built a grade-level passage from the same source, with those specifics intact, plus a key that quotes them.

2. The Letter from Birmingham Jail, reduced to 40 words

Brisk's entire MLK “excerpt” is the single sentence “Injustice anywhere is a threat to justice everywhere…” (about 40 words). The four vocabulary words it asks students to define — *injustice, mutuality, network, destiny* — are simply the nouns in that sentence. RH.9-10.2 asks students to analyze how ideas develop over the course of the text; that cannot be done on one sentence. Diffit built from the whole letter and its key quotes it directly.

3. The math standard names visual models; Brisk has none

5.NF.A.2 reads “by using visual fraction models or equations.” Brisk's worksheet is text-only, tells students to “show your work” while giving a single inline blank, and ships no answer key. Diffit includes bar, circle, and fraction-bar models, a required estimate on each word problem, and a worked key.

4. The inconsistent answer key

Same tool, same Quick Create flow: the Tubman worksheet includes a key (“Examples of Answers”), while the math and MLK worksheets include none. A teacher cannot predict whether a key will appear.

5. A relevel that simplifies the questions but not the reading

Brisk's 2nd-grade relevel cleanly lowers the questions and sample answers — genuinely good work — but “Part 1” stays the adult NPS link, so a second grader is still pointed at an adult website. The relevel cannot lower a reading the worksheet never contained.

What we don't claim

Brisk is not a broken chatbot. It is purpose-built for teachers: its math is accurate, it generates into a clean Google Doc, it accepts a source URL, and it writes well-designed, standards-aware questions. We credit all of that. The Diffit advantage here is not “complete vs. incomplete” at the level of questions — it is whether the lesson is built *around* those questions, with a real source and the standard's full demand.

Brisk did not fabricate. The MLK quotation is real, accurately transcribed, and cited; the Tubman facts are accurate. The divergence is about source-grounding and standards depth, not accuracy.

Brisk's worksheet relevel is genuinely good. The differentiation gap is the reading, not the questions — and we say so.

Three prompts is not a complete evaluation. This covers math, ELA, and HS ELA with one prompt each. Other prompts will exercise other criteria. A district considering both tools should run their own comparison on their own prompts.

This is enablement, not academic research. The rubric is grounded in WestEd's framework and the Diffit Constitution, but scoring was performed by a single evaluator, and the prompts were chosen to exercise the rubric well.

What this comparison shows about the underlying products

Both tools are purpose-built for educators, so the difference is not chat-box-versus-structured-input (the story against a general chatbot). It is about how much of the lesson each tool actually builds.

Brisk optimizes for good questions in a clean document. Diffit optimizes for a complete packet that meets the standard with the real source. Brisk reliably returns thoughtful, standards-aware questions in a tidy Google Doc — a real strength. But it treats the source as a link to cite (or a single quote to lift) and the standard as a label to satisfy with questions, rather than as constraints the whole packet must meet.

Diffit treats the standard and source as inputs that shape the content. The standard's verb drives the activity design (interaction analysis for RI.5.3; visual models for 5.NF.A.2), the source is worked *from* rather than linked, and the answer key is consistently present and grounded. That is why two tools that both write good questions diverge into “a complete, standard-meeting packet” versus “good questions in search of a lesson.”

Methodology — reproducing this comparison

Prompts

Brisk: each artifact was generated with the **Quick Create** flow in the Brisk web app, entering the topic, grade, standard code (selected by hand), and (for the ELA prompts) the source URL, then generating once with no follow-up. Output was produced as a Google Doc and exported to PDF. **Diffit:** topic, grade,

standard (from the Standards picker), and source URL entered as structured inputs, generated once with optional follow-ups skipped.

Scoring

- Diffit and Brisk, June 2026.
- First responses captured before any follow-up prompting. (Brisk was one-shot usable on every prompt; no follow-ups were needed.)
- All artifacts exported as PDF.
- Scoring against rubric.md, one evaluator pass with evidence cited per criterion.
- Per-prompt applicable set = 17 (shared denominator across both columns). N/A in every prompt: §2.4 (Brisk uses no multiple choice) and §4.3 (differentiation, scored in its own section). Math also marks §1.2 N/A (no source). The two ELA prompts also mark §4.4 N/A (a biography and a text-only primary source need no conceptual diagram).
- Diffit's per-criterion column reflects the same Diffit artifacts used in the Diffit vs. Gemini and Diffit vs. MagicSchool comparisons, with one correction made on inspection: Diffit's math error-analysis sheet carries a stray "$" rendering artifact, scored here as a §1.3 Partial (so Diffit's math line is 16/17, not a clean sweep).

Artifacts archived

All artifacts are in artifacts/ (Diffit: diffit-math-fractions, diffit-ela-tubman-v2, diffit-ela-mlk, diffit-ela-tubman-2nd-grade; Brisk: brisk-math-fractions, brisk-tubman-worksheet, brisk-mlk-worksheet, brisk-2nd-grade, plus the supporting brisk-math-resource-pack).

Citation

This comparison applies the Diffit Quality Constitution (drawing on WestEd's research into how teachers judge instructional-materials quality) to a side-by-side artifact comparison between Diffit and Brisk.

Bugler, D., Marple, S., Burr, E., Chen-Gaddini, M., & Finkelstein, N. (2017). *How Teachers Judge the Quality of Instructional Materials: Selecting Instructional Materials, Brief 1 — Quality.* WestEd. https://www.wested.org/resource/selecting-instructional-materials-brief-1-quality/