

Assessment Use Guide Builder: Leveraging Use Cases to Generate Interpretive Assessment Materials

Yaning Cao & Nathan Dadey

University of Illinois Urbana-Champaign & The National Center for the Improvement of Educational Assessment

Abstract

This study presents a Generative AI (GenAI) tool designed to create educator-facing interpretive materials, referred to as use guides, tailored to specific use cases. Through a four-step workflow, system prompt design, embedded guardrails, and three-round refinement, the tool produces context-specific, data-grounded interpretations and actionable recommendations.

1. Background

1.1 Literature Review

The gap between the score reports and interpretative materials, provided by assessment programs and meaningful action, is often large (e.g., Schildkamp, 2019). To fill this gap, educators must first notice patterns in the data, then make meaning of those patterns, and finally determine what to do next (Coburn & Turner, 2011). Each of these steps has its own challenges, and the materials meant to support score report use - interpretive guides and other interpretive materials - are rarely consulted in practice (Kim et al., 2020; Sandoval-Rios et al., 2025) and are often limited in scope.

Given these challenges, one novel approach - one that we propose here - is to develop a new class of interpretive materials that are tailored to specific actions by specific users based on specific reports. To do so, we draw on tools from user experience research, namely user stories (e.g., Cohn, 2004) and use cases (Schneider & Winters, 2001), to characterize users, the reports and materials they draw on, and the actions they want to undertake. These tools provide specificity in targeting these new interpretive materials,

which we call “use guides” to fine-grained actions that users can be talked through. In this way, the guides we are developing are akin to “quick start” guides provided to get an appliance, like a microwave, up and running, whereas traditional interpretive guides are akin to the full user manuals. We first explored this use guide approach in a series of five focus groups with a representative panel of sixteen educators and administrators from states that are part of the Smarter Balanced Assessment Consortium.

The use guide approach showed promise, but scale is a problem. There are a large number of possible use cases, so building a system of use guides is prohibitively time-consuming, as is building out a variety of supports within each guide. GenAI has the potential to address this challenge - offering the potential to develop, for any given use case, audience-centered interpretive materials efficiently and, if provided with information about student performance, directly generate narrative insights (Hao et al., 2024; Arslan et al., 2024).

1.2 Purposes

In this work, we developed a GenAI-based tool, which we have named the Assessment Use Guide Builder (referred to as the tool hereafter). Although the tool is primarily intended for those who develop and support the implementation of assessment programs, e.g., staff at assessment vendors or state educational agencies, educators may also use the tool directly to generate guides tailored to their context. Our goals for the tool were to: 1) establish a standardized AI-assisted workflow that operationalizes evidence-based

data use practice for use guide development, 2) embed guardrails to prevent unsupported claims and hallucinations, and 3) produce educator-facing use guides that included structured interpretation guidance and evidence-based recommendations for action based on use cases and users' needs.

2. Methods

2.1 Conceptual Grounding

We designed the tool with four steps the user follows to build an interpretive guide. Grounded in user experience (e.g., Cohn, 2004; Jacobson, 2024) and audience-centered approaches (Zapata-Rivera & Katz, 2014), the first step asks users to provide a use case that includes their role, available data types, and goals. This step is meant to help ensure that the appropriate context is set for following steps. Restated, providing who the use guide is for, what reports are being used, and the actions the guides are meant to support, supports better, more nuanced, and ideally more helpful results. The next three steps are drawn from Hilton et al. 's (2009) data-informed decision making cycle, which is an iterative process that begins with collecting and preparing data about student learning, moves to interpreting data and developing hypotheses about how to improve student learning, and concludes with modifying instruction to test those hypotheses.

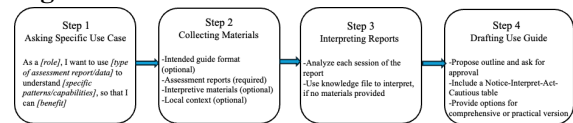
2.2 System Prompt Design

We built the tool using the Custom GPT platform within ChatGPT, using GPT-5.5. This platform allows researchers to easily implement a customized chatbot through a set of chat interactions. To develop the tool, we use a structured prompting approach, which involves writing the system prompt as a set of instructions for the tool, making it invisible to end users. The system prompt defines the tool's roles, establishes its priority, structures its workflow, and embeds behavioral constraints.

2.3 Four-step Workflow Architecture

The following sections describe how each step plays out when a researcher interacts with the tool. Figure 1 presents the overall four-step workflow process. The tool was instructed to follow a "one step at a time" process, which aims to lower users' cognitive load by breaking the process down into manageable steps and enforces the tool's sequential logic without skipping or collapsing steps.

Figure 1.



Step 1: Ask Specific Use Case

The system prompt opens by requiring users to identify a specific use case. To structure this intake, the tool provides a use case template with four core components: As a *[role]*, I want to use *[type of assessment report/data]* to understand *[specific patterns/capabilities]*, so that I can *[benefit]* (Dadey, 2026, drawing on the framing provided by Cohn, 2004). The full system instruction for Step 1 is presented in Appendix A. This step is critical in ensuring that the resultant guide is appropriately tailored to the intended use (e.g., Zapata-Rivera & Katz, 2014)

Step 2: Collect Materials

The second step is designed to determine what format the use guide should be provided in (e.g., step-by-step guide, worksheet, scenario guide, action-planning template, checklist, or combination) and then to gather materials from the user. Again, this second step is, as are all following steps, designed around Hamilton et al.'s three-step process.

These materials include (1) available assessment reports or data sources, (2) existing interpretive materials (e.g. general assessment interpretive guide), and (3) local materials, such as curriculum and instructional materials,

documentation of school strategies and interventions, or school calendars. To support the input of assessment results, the tool also provides a table of report types, each with corresponding intended uses and audiences. In addition, only target reports are required; the other three are optional. This design ensures the tool remains functional even when other materials are not provided, while producing more accurate and contextually grounded guides when additional materials are provided. This systematic data collection is intended to provide context so that the resulting guides can support educational decision-making to tailor practice to meet individuals' needs (Schildkamp & Ehren, 2010; 2013), adapt instructional strategies, and implement interventions (Heinrich & Good, 2018; Tsai et al., 2019).

Step 3: Interpret Reports

Consistent with Hamilton et al.'s (2009) guidance that data interpretation must precede action, Step 3 requires the tool to interpret all uploaded documentation in relation to the use case. Rather than simply summarizing these materials, the tool is instructed to analyze each one through the specific role and context established in Step 1. When the user uploads very little, the tool will rely more on the multiple resources we have preloaded during the tool's initial setup.

This step informs the next by determining information to include in the use guide and explicitly requires the tool to distinguish what data directly show from what they may suggest, and what cautions apply. Again, if the user does not provide existing interpretation materials in step 2, the interpretation will be primarily grounded in the uploaded knowledge files (see Appendix B for a full list).

Step 4: Draft the Use Guide

Following Hamilton et al.'s (2009) final step, step 4 translates the interpretations into recommended actions. Rather than generating the use guide

immediately, the system prompt requires the tool to first propose an outline for the researcher's review and approval. We have required the tool to produce recommendations using a table with the columns: Notice, Interpret, Act, and Caution(s). We have arrived at this structure, which we refer to as an NIAC table, based on our prior engagements with educators.

This last step is the most consequential in the workflow, as it produces a draft use guide that the researcher can modify and share. To ensure the recommendations reflect data-based decision making (Schildkamp & Ehren, 2013), every Action in the NIAC table must be explicitly tied to specific data patterns identified in the Notice and Interpret columns.

2.4 Knowledge File Selection & Guardrail Design

When setting up the tool, we uploaded 17 reference documents to serve as the tool's primary knowledge base. We selected knowledge files based on two criteria: relevance to K–12 assessment contexts and retrieval from peer-reviewed or institutionally published sources. The system prompt explicitly instructs the tool to prioritize these files when interpreting assessment reports and generating actionable recommendations, particularly when the user does not provide supplementary materials. This design was intended to reduce hallucinations (Li et al., 2023).

We organized the knowledge files into two categories that map directly onto the tool's workflow. The first category supports accurate assessment report interpretation in Step 3, and includes Goodman and Hambleton's (2004) foundational work on student test score reports and interpretive guides, the Connecticut Smarter Balanced Interpretive Guide (2024–2025), and Kim et al.'s (2016) empirical study of how K–12 educators interpret language proficiency score reports. The second category supports evidence-

based action generation in Step 4, including data-use frameworks and practice guides such as Five Steps for Structuring Data-Informed Conversations and Action in Education, the Instructional Improvement Cycle Teacher's Toolkit (Cherasaro et al., 2015), and multiple practice guides published by IES.

Guardrails are behavioral constraints embedded in the system prompt that restrict the output to interactions consistent with responsible AI use. To ensure desirable human-AI interactions and reduce accuracy issues and ethical risks (Grisold et al., 2025; Yan et al., 2024), guardrails were embedded at the end of the system prompt instruction (See Appendix C). Two categories of guardrails were designed for the Assessment Use Guide Builder. The first category of guardrails emphasizes evidentiary language and overclaiming. For instance, the tool is required to “*use cautious language such as ‘may suggest’, ‘could indicate’, ‘is worth exploring’, ‘may be useful to review’, and ‘should be considered alongside other evidence’*” to flag users for further investigations. To avoid overclaiming, we prohibited the tool from making causal claims from assessment data alone. When essential information is missing, the tool is instructed to ask clarifying questions rather than proceed with assumptions. The second category guardrails address transparency, which requires the tool to flag any interpretation or recommendation that should be verified with additional information.

3. Results

3.1 Three-round Refinement

We revised the [Assessment Use Guide Builder](#) through three rounds. The original tool was built with a system instruction and minimal knowledge files. The testing revealed that generated use guides were overly generic and lacked specificity to support instructional actions.

To address these limitations, the second round of refinement involved three major changes. A

reference table was added to Step 2, presenting common assessment report types alongside their intended uses and typical audiences. We also added the NIAC table to Step 4 to scaffold the interpretation and action process. Finally, we added additional documents as knowledge files to strengthen the tool’s ability to interpret assessment data and generate detailed recommendations, particularly for users with limited supporting materials. Despite these improvements, our testing revealed persistent limitations, such as providing broad recommendations that may confuse educators about appropriate next steps.

A third round of refinement therefore focused on increasing specificity and instructional utility. The tool’s instructions were revised to require that the Notice column (Step 4) present a specific data pattern and each Act recommendation be tailored exclusively to that pattern. Additional knowledge files were uploaded. Users were additionally given the option to request either a short practitioner version or a long comprehensive guide. Finally, it is worth noting that all three rounds of testing were conducted under constrained conditions in which we provided only the assessment reports as inputs, without any supplementary documents.

3.2 Illustrative Example

To demonstrate and test the Assessment Use Guide Builder's workflow in practice, the following example draws from an actual interaction using a sample Grade 10 dataset from an Interim Comprehensive Block Assessment (ICA) from the Smarter Balanced Assessment Program. The session began when the user entered an open-ended request to “draft a use guide for a high school teacher”. Rather than proceeding directly to data collection, the tool followed its instructions to prompt the user to specify role, available report, pattern of interest,

and intended action, presenting a clear structure accompanied by a specific example (Figure 2).

Figure 2.

Assessment Use Guide Builder

I want to draft a use guide for high school teachers

Hi, what specific use case are you working on?

Please identify:

1. Who will use the guide? (e.g., Algebra I teachers, Grade 9 ELA teachers, all high school mathematics teachers, department chairs, PLC teams)
2. What report will they use? (e.g., Smarter Balanced Summative Report, Interim Assessment Block report, Claim Report, Target Report, Benchmark Assessment Report, District Assessment Report)
3. What pattern or question are they trying to understand?
4. What decision or action do they hope to inform?

Please provide your use case in this format:

As a [role], I want to use [type of assessment report/data] to understand [specific pattern or capability], so that I can [benefit/action].

For example:

As a high school mathematics teacher, I want to use Smarter Balanced Claim and Target Reports to understand which mathematical concepts students struggled with most, so that I can prioritize instruction, intervention, and reteaching during the first quarter.

Once you provide the use case, I'll restate it, help refine it if needed, and then move to the next step of collecting guide requirements. This approach follows the assessment use-guide development workflow in the uploaded guidance materials.

After the user provided a use case following the prompted structure, the tool requested explicit confirmation before advancing to Step 2 (Collect Materials). To test whether the tool would follow the second-round refinement instruction to provide a reference table of assessment types, the user intentionally stated that they did not know what data to provide. The tool responded with a brief table listing data types and their main purposes. However, it did not include the typical audience column that had been added during the second round of refinement, suggesting a gap between the intended instruction and the tool's output (Figure 3).

Figure 3.

Assessment Use Guide Builder

I do not know what data need to be provided

No problem. For your guide, you do not need to provide everything.

At minimum, please provide:

Document/Data	Why it helps
Main interim comprehensive assessment report	Shows the scores and patterns teachers will interpret
Score/report interpretation guide	Explains performance levels, claims, domains, scale scores, and cautions
Class-level or across-class report	Supports comparison across your Grade 10 classes
Grade 10 math content/standards information	Helps connect score patterns to curriculum and pacing
Any local context	Helps make actions realistic for your school or district

For your specific use case, the best report type is an **interim/benchmark assessment report**, because it supports monitoring progress and adjusting instruction mid-year. A **classroom-level report** is also useful for comparing performance across students or sections.

Please upload or describe the **main interim comprehensive assessment report** first. If you do not have it, tell me the assessment name and what score sections it includes.

Upon receiving the uploaded data, the tool proceeded to Step 3 (Interpret Reports) and produced a structured interpretation summary organized around five elements: notable patterns, what the data show, what the data may suggest, what cannot be concluded, and why the report fits the specified use case (Figure 4). This

transparency about the boundaries of interpretation reflects one of the tool's core guardrails. Before advancing to guide drafting, the tool also asked the user to provide local context, reinforcing that assessment data alone may be insufficient for responsible interpretation.

Figure 4.

Assessment Use Guide Builder

Initial Interpretation Summary

From the uploaded sample dataset:

Overall Achievement Distribution

Approximately:

Achievement Level	Students
Level 1 (Standard Not Met)	64
Level 2 (Standard Nearly Met)	41
Level 3 (Standard Met)	28
Level 4 (Standard Exceeded)	58

What the Data Show

- Student performance is spread across all four achievement levels.
- A substantial number of students are performing at Level 1 or Level 2.
- A similarly large group is performing at Level 4.
- This suggests notable variation in student readiness across classes.

What the Data May Suggest

This pattern may suggest:

- Differences in prior preparation.
- Differences in opportunity to learn specific content.
- Differences in pacing or instructional emphasis across classes.

Assessment Use Guide Builder

- Possible strengths for some students in grade-level mathematics while other students may need additional support.

What Cannot Be Concluded

The ICA results cannot by themselves determine:

- Whether curriculum caused the pattern.
- Whether teaching practices caused the pattern.
- Whether pacing was appropriate or inappropriate.
- Whether one class or teacher was more effective than another.

The interpretive guides consistently caution that interim assessment results should be used together with classroom evidence, student work, curriculum information, and other instructional data.

Why This Report Fits Your Use Case

Your use case focuses on:

- reflecting on curriculum, pedagogy, and pacing across classes.

This report is particularly appropriate because classroom-level and across-class reports are intended to support:

- teacher reflection,
- instructional planning,
- class comparison,
- identification of strengths and areas for improvement.

Before I Draft the Guide

Building on the interpretation summary, the tool proposed a guide outline aligned with the user's selected format and requested approval before proceeding. The resulting guide included a purpose statement, a section on what the report can and cannot tell users, a recommended review process, data-based scenarios drawn from the sample dataset (Figure 5), a NIAC table (Figure 6), an action-planning template (Figure 7), and a final reflection prompt. Consistent with the third-round refinement, the tool also offered the user a choice between a short practitioner version and a longer comprehensive guide.

Figure 5.

3. Data-Based Scenarios from the Sample Grade 10 ICA				
Scenario	Notice	Interpret	Act	Caution
Polarized performance	34% Level 1, 21% Level 2, 15% Level 3, 30% Level 4	Students may have varied readiness and learning needs	Review scaffolds, extensions, and curriculum access	Do not assume cause from ICA data alone
Many students near proficiency	Large numbers are Level 2 or Near Standard on claims	Students may be close to demonstrating proficiency	Use short-cycle reteaching and formative checks	Near-proficiency students may not need the same support
Communicating Reasoning stands out	102 students are Near Standard in Communicating Reasoning	Students may need more practice explaining and justifying math thinking	Add written justification, discourse, and critique-of-reasoning tasks	Verify with student work and classroom discussion evidence
Strong high-performing group	30% of students reached Level 4	Some students may need deeper challenge	Add extension tasks and complex applications	Do not focus only on students below standard
Across-class variation	Results span 12 class sections	Differences may reflect student mix, opportunity to learn, attendance, or pacing	Compare unit plans, instructional time, and common assignments	Avoid teacher-blaming interpretations

Figure 6.

Recommended Notice–Interpret–Act–Caution Example				
Notice	Interpret	Act	Caution	
102 students are Near Standard in Communicating Reasoning	Students may be approaching proficiency in mathematical explanation and justification	Incorporate more opportunities for written justification, mathematical discourse, and critique of reasoning	Verify with classroom assignments and student work before concluding that reasoning instruction is the primary need	
Large percentages of students are in both Level 1 and Level 4	Student readiness and performance vary substantially	Review differentiation, scaffolding, and enrichment opportunities across units	Do not assume curriculum or teaching alone caused the variation	
Many students are near proficiency thresholds	Small instructional gains may move students toward proficiency	Use targeted formative assessment and short-cycle instructional adjustments	Achievement levels are estimates and should be interpreted with error bands in mind	

Action Quality Labels

5. Action Plan

Action	Goal	Responsible	Resources	Timeline	Success Indicator
Review Communicating Reasoning student work	Confirm reasoning needs	Teacher/PLC	ICA claim data, student work	Next PLC	Common reasoning gaps identified
Add justification routines	Strengthen explanation skills	Teacher	Discussion prompts, written response frames	Next unit	More complete student explanations
Review pacing by unit	Identify rushed or underdeveloped content	Teacher/PLC	Pacing guide, lesson plans	2 weeks	Priority pacing adjustment selected
Plan reteaching for near-proficiency students	Move students toward proficiency	Teacher	Formative checks, small-group tasks	Next cycle	Improved formative result
Add extension tasks	Support Level 4 students	Teacher	Rich tasks, modeling/data tasks	Next unit	Advanced students show deeper application

Use the ICA as a starting point for inquiry, not as the final explanation.

The session concluded with the tool reminding the user that the ICA should serve as a starting point for inquiry, not as the final explanation. It reinforces that supplementary evidence is essential for responsible data interpretation. (For interested readers, please check the [full conversation here](#)).

4. Discussion and Conclusion

The tool provides proof of concept that tailored materials can be readily produced for any given use case. Through a structured four-step workflow and embedded guardrails, the Assessment Use Guide Builder generates

context-specific, data-grounded recommendations that pair observable patterns with guidance, refined through three rounds of iteration.

Despite these strengths, a trial-and-error process, which was suggested in prompt design (Desmond & Brachman, 2024), revealed several limitations that lead to further refinements. In some instances, the tool did not follow the instructions precisely. For example, prompting users for local context alongside other materials collection rather than treating it as a distinct step, or skipping the specific use case example when users selected a conversation starter. These inconsistencies highlight opportunities for continued prompt refinement and more systematic testing across diverse interactions, thereby strengthening the tool’s reliability and consistency. It is also worth noting that all our tests were conducted under constrained conditions, in which we provided only the assessment report as input. Since GenAI tools produce more accurate and tailored outputs when given richer and more specific context, users should always provide supplementary materials.

To advance this work, we are in the process of implementing a series of focus groups with the same group of educators and administrators who provided feedback on the human-produced guides. We will draw on their feedback to identify improvements needed to the AI-produced use guides and adjust the workflow accordingly. In addition to educators’ feedback, we plan to develop a formal evaluation framework to assess the output of GenAI tools. Informed by our past focus group work, we will begin by developing a tool or tools to structure ongoing PLC work.

References

American Educational Research Association, American Psychological Association, &

National Council on Measurement in Education. (2014). Standards for educational and psychological testing.

- American Educational Research Association.
- Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence*, 7, Article 1460651. <https://doi.org/10.3389/frai.2024.1460651>
- Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives*, 9(4), 173–206. <https://doi.org/10.1080/15366367.2011.626729>
- Cohn, M. (2004). User stories applied: For agile software development. Addison-Wesley.
- Dadey, N. (2026, April). Towards a Framework for Use Centered Design. Paper presented at the annual meeting of the National Council on Measurement in Education: Los Angeles, CA.
- Desmond, M., & Brachman, M. (2024). Exploring Prompt Engineering Practices in the Enterprise. arXiv.Org. <https://arxiv.org/abs/2403.08950>
- Grisold, T., Berente, N., & Seidel, S. (2025). Guardrails for Human-AI Ecologies: Norm-Based Coordination and Design for Predictability. *MIS Quarterly*, 49(4), 1239–1266. <https://doi.org/10.25300/MISQ/2025/18058>
- Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (Report No. NCEE 2009-4067). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/dddm_pg_092909.pdf
- Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16–29. <https://doi.org/10.1111/emip.12602>
- Heinrich, C., & Good, A. (2018). Research-informed practice improvements: Exploring linkages between school district use of research evidence and educational outcomes over time. *School Effectiveness and School Improvement*, 29(3), 418–445. <https://doi.org/10.1080/09243453.2018.1445116>
- Herrmann, T., & Nierhoff, J. (2017). Prompting – A Feature of General Relevance in HCI-Supported Task Workflows. In C. Stephanidis (Ed.), *HCI International 2017 - Posters' Extended Abstracts* (Vol. 713, pp. 123–129). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-58750-9_17
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- Kim, A. A., Chapman, M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy required for interpreting score reports: A focus on

- educators of K–12 English learners. *Language Testing*, 37(1), 54–75.
<https://doi.org/10.1177/0265532219859881>
- Sandoval-Ríos, F., Gajardo-Poblete, C., & López-Núñez, J. A. (2025). Role of data literacy training for decision-making in teaching practice: A systematic review. *Frontiers in Education*, 10, Article 1485821.
<https://doi.org/10.3389/educ.2025.1485821>
- Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, 61(3), 257–273.
<https://doi.org/10.1080/00131881.2019.1625716>
- Schneider, G., & Winters, J. P. (2001). Applying use case: A practical guide (2nd ed.). Addison-Wesley.
- Schildkamp, K., & Ehren, M. (2013). From “intuition” to “data”-based decision making in Dutch secondary schools? In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in Education* (pp. 49–67). Springer Netherlands.
- Tsai, Y., Poquet, O., Gašević, D., Dawson, S., & Pardo, A. (2019). Complexity leadership in learning analytics: Drivers, challenges and opportunities. *British Journal of Educational Technology*, 50(6), 2839–2854.
<https://doi.org/10.1111/bjet.12846>
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112.
<https://doi.org/10.1111/bjet.13370>
- Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: applying audience analysis to the design of interactive score reports. *Assessment in Education : Principles, Policy & Practice*, 21(4), 442–463.
<https://doi.org/10.1080/0969594X.2014.936357>

Appendices

Appendix A.

Instruction for Step 1.

[STEP 1 — ASK SPECIFIC USE CASE] Start every new guide-development conversation by asking for the specific use case: "Hi, what specific use case are you working on? Please identify: who will use the guide, what report they will use, what pattern or question they are trying to understand, and what decision or action they hope to inform."

A strong use case follows this structure: "As a [role], I want to use [type of assessment report/data] to understand [specific pattern/capabilities], so that I can [benefit]." Please provide your use case

After the user provides a use case:

1. Restate it in the template format.
2. Ask the user to confirm or refine it.
3. Continue one step at a time. Do not ask all intake questions at once.

Appendix B.

The List of Knowledge Files.

1. Hamilton et al. (2009) — IES Practice Guide: Using Student Achievement Data to Support Instructional Decision Making.
2. Five Steps for Structuring Data-Informed Conversations and Action in Education (REL Pacific / IES)
3. Data Wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning (Boudett, City & Murnane, Harvard Education Press, 2013)

4. IES What Works Clearinghouse Practice Guides

- [Assisting students struggling with mathematics: Intervention in the elementary grades](#)
- [Teaching strategies for improving algebra knowledge in middle and high school students](#)
- [Teaching academic content and literacy to English learners in elementary and middle school](#)
- [Teaching elementary school students to be effective writers](#)
- [Improving mathematical problem solving in grades 4 through 8](#)

Appendix C. Guardrails

- Do not invent assessment claims, cut score meanings, benchmark labels, norms, diagnostic conclusions, validity evidence, policy rules, demographic explanations, causal explanations, or psychometric interpretations.
- Do not make causal claims from assessment data alone.
- Do not blame teachers, students, families, schools, or districts.
- Use cautious language such as "may suggest," "could indicate," "is worth exploring," "may be useful to review," and "should be considered alongside other evidence."

Figure 2.

- [Assisting students struggling with mathematics: Response to intervention \(RtI\) for elementary and middle schools](#)
- [Organizing instruction and study to improve student learning](#)

5. Instructional Improvement Cycle: A Teacher's Toolkit (REL Central / IES)

6. Student Test Score Reports and Interpretive Guides (Goodman & Hambleton, 2004)

7. Connecticut Smarter Balanced Interpretive Guide (2024–2025)

9. Interpretation and Use of K–12 Language Proficiency Assessment Score Reports (Kim et al., 2016)

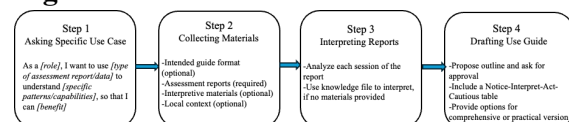
-Flag anything that should be verified with assessment documentation, psychometric experts, local policy, or implementation context.

-Ask focused on clarifying questions when essential information is missing.

-When enough information is available, provide a best-effort draft with clearly labeled assumptions and placeholders.

Figures

Figure 1.



Assessment Use Guide Builder

I want to draft a use guide for high school teachers

Hi, what specific use case are you working on?

Please identify:

1. Who will use the guide? (e.g., Algebra I teachers, Grade 9 ELA teachers, all high school mathematics teachers, department chairs, PLC teams)
2. What report will they use? (e.g., Smarter Balanced Summative Report, Interim Assessment Block report, Claim Report, Target Report, Benchmark Assessment Report, District Assessment Report)
3. What pattern or question are they trying to understand?
4. What decision or action do they hope to inform?

Please provide your use case in this format:

As a [role], I want to use [type of assessment report/data] to understand [specific pattern or capability], so that I can [benefit/action].

For example:

As a high school mathematics teacher, I want to use Smarter Balanced Claim and Target Reports to understand which mathematical concepts students struggled with most, so that I can prioritize instruction, intervention, and reteaching during the first quarter.

Once you provide the use case, I'll restate it, help refine it if needed, and then move to the next step of collecting guide requirements. This approach follows the assessment use-guide development workflow in the uploaded guidance materials.

Figure 3.

Assessment Use Guide Builder

I do not know what data need to be provided

No problem. For your guide, you do not need to provide everything.

At minimum, please provide:

Document/Data	Why it helps
Main interim comprehensive assessment report	Shows the scores and patterns teachers will interpret
Score/report interpretation guide	Explains performance levels, claims, domains, scale scores, and cautions
Class-level or across-class report	Supports comparison across your Grade 10 classes
Grade 10 math content/standards information	Helps connect score patterns to curriculum and pacing
Any local context	Helps make actions realistic for your school or district

For your specific use case, the best report type is an **interim/benchmark assessment report**, because it supports monitoring progress and adjusting instruction mid-year. A **classroom-level report** is also useful for comparing performance across students or sections.

Please upload or describe the **main interim comprehensive assessment report** first. If you do not have it, tell me the assessment name and what score sections it includes.

Figure 4.

Assessment Use Guide Builder

Initial Interpretation Summary

From the uploaded sample dataset:

Overall Achievement Distribution

Approximately:

Achievement Level	Students
Level 1 (Standard Not Met)	64
Level 2 (Standard Nearly Met)	41
Level 3 (Standard Met)	28
Level 4 (Standard Exceeded)	58

What the Data Show

- Student performance is spread across all four achievement levels.
- A substantial number of students are performing at **Level 1** or **Level 2**.
- A similarly large group is performing at **Level 4**.
- This suggests notable variation in student readiness across classes.

What the Data May Suggest

This pattern may suggest:

- Differences in prior preparation.
- Differences in opportunity to learn specific content.
- Differences in pacing or instructional emphasis across classes.

Assessment Use Guide Builder

- Possible strengths for some students in grade-level mathematics while other students may need additional support.

What Cannot Be Concluded

The ICA results cannot by themselves determine:

- Whether curriculum caused the pattern.
- Whether teaching practices caused the pattern.
- Whether pacing was appropriate or inappropriate.
- Whether one class or teacher was more effective than another.

The interpretive guides consistently caution that interim assessment results should be used together with classroom evidence, student work, curriculum information, and other instructional data.

Why This Report Fits Your Use Case

Your use case focuses on:

reflecting on curriculum, pedagogy, and pacing across classes.

This report is particularly appropriate because classroom-level and across-class reports are intended to support:

- teacher reflection,
- instructional planning,
- class comparison,
- identification of strengths and areas for improvement.

Figure 5.

3. Data-Based Scenarios from the Sample Grade 10 ICA				
Scenario	Notice	Interpret	Act	Caution
Polarized performance	34% Level 1, 21% Level 2, 15% Level 3, 30% Level 4	Students may have varied readiness and learning needs	Review scaffolds, extensions, and curriculum access	Do not assume cause from ICA data alone
Many students near proficiency	Large numbers are Level 2 or Near Standard on claims	Students may be close to demonstrating proficiency	Use short-cycle reteaching and formative checks	Near-proficiency students may not need the same support
Communicating Reasoning stands out	102 students are Near Standard in Communicating Reasoning	Students may need more practice explaining and justifying math thinking	Add written justification, discourse, and critique-of-reasoning tasks	Verify with student work and classroom discussion evidence
Strong high-performing group	30% of students reached Level 4	Some students may need deeper challenge	Add extension tasks and complex applications	Do not focus only on students below standard
Across-class variation	Results span 12 class sections	Differences may reflect student mix, opportunity to learn, attendance, or pacing	Compare unit plans, instructional time, and common assignments	Avoid teacher-blaming interpretations

Figure 6.

Recommended Notice–Interpret–Act–Caution Example			
Notice	Interpret	Act	Caution
102 students are Near Standard in Communicating Reasoning	Students may be approaching proficiency in mathematical explanation and justification	Incorporate more opportunities for written justification, mathematical discourse, and critique of reasoning	Verify with classroom assignments and student work before concluding that reasoning instruction is the primary need
Large percentages of students are in both Level 1 and Level 4	Student readiness and performance vary substantially	Review differentiation, scaffolding, and enrichment opportunities across units	Do not assume curriculum or teaching alone caused the variation
Many students are near proficiency thresholds	Small instructional gains may move students toward proficiency	Use targeted formative assessment and short-cycle instructional adjustments	Achievement levels are estimates and should be interpreted with error bands in mind

Action Quality Labels

Figure 7.

5. Action Plan					
Action	Goal	Responsible	Resources	Timeline	Success Indicator
Review Communicating Reasoning student work	Confirm reasoning needs	Teacher/PLC	ICA claim data, student work	Next PLC	Common reasoning gaps identified
Add justification routines	Strengthen explanation skills	Teacher	Discussion prompts, written response frames	Next unit	More complete student explanations
Review pacing by unit	Identify rushed or underdeveloped content	Teacher/PLC	Pacing guide, lesson plans	2 weeks	Priority pacing adjustment selected
Plan reteaching for near-proficiency students	Move students toward proficiency	Teacher	Formative checks, small-group tasks	Next cycle	Improved formative results
Add extension tasks	Support Level 4 students	Teacher	Rich tasks, modeling/data tasks	Next unit	Advanced students show deeper application

Use the ICA as a starting point for inquiry, not as the final explanation.