RYAN K. BOETTGER

WHITE PAPER

Rubrics That Hold Up: Using AI to Draft, Stress-Test, and Maintain Assessment

Ryan K. Boettger · June 12, 2026 · White Paper

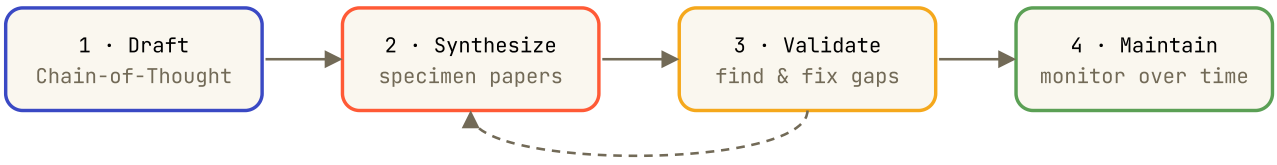
> Drawn from my workshop *Innovative Assessment Strategies in the Age of AI*, delivered for the Texas Higher Education Coordinating Board's AI Facilitated Learning Network. The talk page has the full session materials.

Let's start with the uncomfortable truth: most rubrics are never actually tested. We write them the night before an assignment goes out, we use them to assign grades that affect real students, and we almost never check whether they measure what we think they measure. We just trust them.

I've spent a long time on that problem — and not abstractly. Back in 2010 I published a study in *IEEE Transactions on Professional Communication* on the unglamorous work of building rubrics that are actually **valid** (they target the purpose you built them for) and **reliable** (they score consistently, no matter who's holding the pen). The single most important lever I found for reliability was something called **specimen papers**: a set of real student samples, one for each level of your scale, that you use to calibrate scoring before you grade anything that counts.

Here's the catch that's haunted that method for fifteen years — *good specimen papers are expensive.* You collect them over semesters, or you hand-pick them and hope they cover the range. I ended that 2010 paper predicting that assessment would have to evolve with technology. It did. And the most useful thing AI does for assessment isn't writing your rubric — it's **manufacturing specimen papers on demand so you can finally stress-test the thing before a student pays for its blind spots.**

The workflow



Four steps — and that dashed loop between 2 and 3 is where the real work happens.

1 · Draft with Chain-of-Thought prompting

Don't ask a model to "make a rubric" cold. Walk it through its thinking first — prime it, *then* give it the assignment. You'll get noticeably sharper criteria:

```
I'm a college professor. What can you tell me about assessment rubrics?
```

```
Below is an assignment I give in my class. Please generate an analytic
rubric for it.

[Assignment]
…your assignment here…
[/Assignment]
```

What you have now is a *draft*. Treat it as exactly that — because we haven't tested it yet.

2 · Synthesize specimen papers — the step everyone skips

This is the move that turns AI from a shortcut into a genuine upgrade on the 2010 method. Before you trust the rubric, ask the model to write **synthetic student responses** to the same assignment — an excellent one, a competent one, a weak one, and a sneaky one that follows the directions but misses the point entirely:

```
Write four responses to this assignment at different quality levels:
one excellent, one competent, one weak, and one that technically
follows the instructions but misses the point. Vary them realistically.
```

That's a set of specimen papers — the very thing I used to spend three semesters collecting — generated in about thirty seconds, and you can dial in any edge case you're worried about.

3 · Validate and refine

Now score those specimens with your draft rubric, by hand or by asking the model to apply it. **The disagreements are the gold.** When the rubric rates the hollow response highly, or can't tell "competent" from "excellent," you've found a validity problem — the kind that, in 2010, only surfaced after real grades were already in the gradebook. Fix the criteria, sharpen the descriptors, adjust the weights, and run the specimens again:

```
Adjust the rubric to weight clarity and problem-solving more heavily,
and rewrite the descriptors so "competent" and "excellent" are clearly
distinguishable.
```

One caution from the old research that's only more true now: every rater brings biases, and **AI is a rater too.** It has its own. So check whether *it* scores consistently, the same way you'd check inter-rater reliability between two human graders — and don't outsource the final call.

4 · Maintain and monitor

Assignments drift, cohorts change, and a rubric that was valid in 2024 may quietly stop being valid by 2026. The cost of *re-validating* used to be prohibitive, so we mostly didn't. Now it isn't — regenerate a fresh batch of specimens now and then, re-score, and watch for the rubric starting to slip. Maintenance becomes a habit instead of a rewrite.

Beyond rubrics

Rubrics aren't the only instrument, and they're not always the right one. The same loop works for **objective assessments** — multiple-choice, true/false, matching — generated from your course materials and then stress-tested against synthetic answers for ambiguity and giveaway patterns:

```
Below are course materials I use in my class. Please generate
multiple-choice questions based on this material, then flag any item
where a test-wise student could guess the answer without knowing it.
```

Where do you stop?

The hardest question from the workshop wasn't technical. It was this: *as you let AI into rubric-writing, specimen generation, scoring, and monitoring — where do you personally feel comfortable stopping?*

There's no universal line, but a few principles travel well. **Synthetic specimens are a stress test, not a substitute** for real student work; they expose a rubric's blind spots, they don't certify its fairness. **A human owns the judgment** — AI drafts and challenges the instrument, but a person decides what the assessment values and signs off on the grade. And **be transparent**: students deserve to know the rubric, and increasingly, how it was built.

Used this way, AI doesn't replace the work of assessment. It finally makes the part we always skipped — checking whether the rubric holds up — cheap enough to actually do.