🚨 NEGATIVE RUBRIC ITEMS RECAP 🚨

Let’s suppose that our prompt *explicitly* asks for the following output:

- Cohen's d effect size for Age between survivors and non-survivors

Let’s assume the following is a *fact*:

- Cohen's d effect size for Age between survivors and non-survivors is -0.157.

So, we need a *positive* rubric item stating:

- **PC1**
  “Reports Cohen's d effect size for Age between survivors and non-survivors as -0.157.” [35]

PC1 is a **Must-Have** criterion because the prompt explicitly asked for such an output.

We also need negative rubric items:

- ➔ *Can we add a negative rubric item regarding Cohen's d effect size for the Age Difference between survivors and non-survivors?*
  - ◆ The **ONLY** valid scenarios where we can add a negative criterion for this output are those wherein:
    - The model *actually* makes a mistake in providing the correct value for Cohen's d effect size, for example, -2.789 (which is incorrect).
    - If applicable, the negative rubric item for that example should be written like this:
      - ○ **NC1:**
        “Reports Cohen's d effect size for Age between survivors and non-survivors as -2.789.”

        - As you can see, **NC1** is *almost the same as* the positive item, **PC1**, **BUT** the particular value that **NC1** specifies is:
            - The value provided by the GPT-5, and,
            - Such a value is incorrect (the correct value is -0.157, but the model stated that it is -2.789).

- Can we write a negative rubric item referring to Cohen's d effect size by using the *generalized wording*, as in “Reports that X is other than Y”?
    - **NO, WE CANNOT**: because *generalized wording* is reserved for **Nice-To-Have** criteria **exclusively**. As we noted, Cohen's d effect size at issue is *an output the prompt explicitly asked for*, that is, a **Must-Have**. So, a criterion like the example below would be an ***invalid negative rubric item*:**

        - **NC2**
          “Reports that Cohen's d effect size for Age between survivors and non-survivors *is other than* -0.157” [-35]

        - Criterion **NC2** *overlaps* with **PC1** because adding **NC2** to the rubric makes the rubric penalize the model twice in an **INVALID** way:
            - First, GPT-5 doesn’t win the 35 points awarded by **PC1**, given that GPT-5 stated that Cohen’s d effect size at stake is -2.789.
            - Second, GPT-5 loses 35 additional points, given that GPT-5 stated that Cohen’s d effect size at stake is other than -0.157.

The only valid scenarios wherein we can use *generalized wording* (i.e., “Reports that X is other than Y”) are those meeting this exact condition:

- The negative rubric item concerns a Nice-to-Have output, that is, something that the prompt didn’t explicitly ask for.

How to write a Nice-to-Have Negative Rubric Item?

Negative rubric items must use the wording “other than Y” to cover ALL possibly incorrect answers that GPT-5 *could* provide, that is, all the answers where the value stated by GPT-5, say X, is not Y.

To summarize, let’s take a look at the following table covering all the possible types of Negative Rubric Items, both Valid and Invalid ones, to realize the contrast between them:

Symbol “✅” indicates the combination is valid; symbol “❌” indicates the combination is *invalid*.

	Negative	
	Nice to Have	Must Have
Generalized	✅ (A)	❌ (B)
Particularized	❌ (C)	✅ (D)

Examples of each possible combination of Negative Rubrics

✅ **A (Valid, Nice-to-Have)** = "States that the standard deviation is any other than 9.5", *while*

(1) The actual standard deviation is 9.5, &
(2) The prompt **DIDN’T** require stating the standard deviation.

It’s valid because it covers an incorrect nice-to-have claim, and does so in a general way (all the possibly wrong values).

❌ **B (Invalid, Must-Have)** = "States that the mean age for women is any other than 20", *while*

(1) The mean age for women is 20, &
(2) The prompt **DID** require stating the mean age for women.

It's invalid because it covers an incorrect must-have claim, but in a general way.

❌ **C (Invalid, Nice-to-Have)** = "States that the standard deviation is 1.8", *while*
(1) The actual standard deviation is 9.5, &
(2) The prompt **DIDN'T** require stating the standard deviation.

It's invalid because it covers an incorrect nice-to-have claim, but in a specific way.

✅ **D (Valid, Must-Have)** = "States that the standard deviation is 7", *while*
(1) The actual standard deviation is 9.5, &
(2) The prompt DID require stating the standard deviation.

It's valid because it covers an incorrect must-have claim, and does so in a specific way.

REMEMBER: This type of negative rubric item is valid ONLY when GPT-5 actually committed the error (here, claiming that the standard deviation is 7).