**Example Feedback on Interaction Metrics Projects**

Feedback on presentations was given immediately after the presentation but *no points were considered*. The students could use this feedback to improve their final report, which was due 3-5 days later. Feedback that the students received is in **purple** ink. After three examples of presentation feedback, there is report feedback for the same teams.

**Interaction Metrics Feedback on Presentation**

***Helicopter Team (analyzing communication within helicopter crews)***

Stephen’s note for colleagues: This team was in real trouble at presentation time and was missing several components. I met with the team to see what they could do to save their grade, and then their final project was much better, but not optimal.

It would have been nice in the front of the presentation to tell me what your goal was in terms of metrics. What were you trying to measure in the helicopter? See if you can include that in the report, so on the very first page of text I know what you’re trying to accomplish with metrics.

There was too much background, really. In your video, at the 8:00 minute point, more than half-way through, you’re still discussing background. In the report, do include what you’ve done for the background (esp. how it relates to what you did), but what I really want to learn about is what you did.

Good slide design; not too crowded with text, and messages of each slide were clear.

I’d be interested in whether there were any methods or specific factors in choosing the audio recordings that you chose.

What were your 3 metrics? The assignment notes that you need to explore 3 metrics.

I like the limitations section. You might extend that in the report from what you didn’t have to what you think you could accomplish if you did have that missing info.

Overall, the presentation didn’t give me the impression that you did a lot of work. It’s supposed to be a major project, but the results in the slides are just one table. From my Q&A with [team member] in class, it sounds like you did do a lot of work, but maybe it just didn’t come through in the presentation. Please make that clear in the report.

[Team member] talked about not having as much data as you wanted, which made it hard to develop metrics. That could be ok, but then treat your data as pilot data that would help you design metrics that you could use if you did have more information.

Please read the instructions in the assignment carefully. There need to be annotated event logs.

Currently based on the presentation, I’m worried this project wouldn’t earn a passing grade, though I haven’t run the numbers. If you need more time than my current deadline for the report, we can discuss that.

***Design Studio Team (creating metrics for design classes that use studio / critique model)***

Stephen’s note for colleagues: this team had a designer on it, as well as two non-designers. This can be more difficult to grade in comparison to other teams because everything produced by designers looks far more attractive than anyone else’s presentation, but you still need to analyze the content.

Great brief slides on the background research on feedback. In your report, see if you can tie your results back to that background research at the end, e.g., “Our results align with… but conflict with …”

Beautiful slide design. Had more content per slide, but it’s just right; not too crowded.

I like the image / flowchart of critique flow. Is that innovative in some way? Could it lead to a paper? Or maybe not, if design instructors react with “Yeah, duh.”

There’s generally a principle of “whenever you tell me a mean, also tell me the stdev” (or some form of variance). For your bar charts, please add error bars for std dev if those are means.

I really like your analysis of the table seating arrangements!

Be sure to include all your presentation imagery in the report as well.

Results slide with table of green is a little too green in some cells (green text on green background is hard to read).

On Results slide with Excel Regression table data, might change the precision to 2 decimal places for all numbers.

The word “this”! [Student] described its prevalence in design critiques well, but the word “this” didn’t appear on a slide as something important.

On the Discussion slide with the 2x2 matrix, I was wishing for some data from your data collection on there. Or maybe you start with that slide and then add some data to it, even if it’s just examples of words spoken by your participants.

When you talk about people feeling shy about being recorded, you might cite the Hawthorne effect.

Good future directions list!

***Combat Team (creating metrics to measure communication within both real Army patrols and military multiplayer videogames)***

Stephen’s note for colleagues: this team did clever work creating metrics for doing behavioral coding (counting specific behavioral markers) in YouTube video recordings of both live Army combat and of multiplayer combat games. By the way, IRR = Interrater Reliability.

Nice slide design.

I like the metrics you explored. I wondered if you had any data from the video game that could be used to validate them, e.g., our metrics predicted 94% for performance, and indeed this team went on to win multiple other games online, which demonstrates their high competence.

Somehow in the presentation (maybe I was taking notes?) I missed what 1/4, 2/4, etc. meant.

In the results, watch out for having more precision that really makes sense, e.g., 93.89%. Usually, the precision of resulting number from math should have just 1 more decimal place than the numbers that were used as inputs to the calculation. (And there are diff rules than that for engineers.)

Nice analysis of IRR. You could make the point that ease of having high IRR is a criterion for judging a metric as well as its ability to predict the behavior, etc.

Great leveraging of the Salas et al. book’s Table 21.2 in Discussion.

Good suggestions for improvement. Those could be the seeds of research papers!

It was clear from the presentation that you have done a lot of work. Be sure to make that clear in your report, noting what’s your original work with metric design vs. your using existing metrics.

**Interaction Metrics Feedback on Reports**

This feedback was given to teams after they turned in their final reports. Point deductions were smaller than I do in some classes because this project was 35%of their whole course grade.

***Helicopter Team***

**33/35 points**

**Comments on report**

Overall, the writing in this report is very strong.

Good job including appropriate citations in the Introduction.

Nice definitions of your 4 metrics in the Introduction.

Your metrics are interesting and more complex because more is not always better. Your flight crew team needs the right amount of communication, the right amount of call and response, not too much and not too little.

You didn’t have a lot of time, but if you were going to write such a report in the future, I’d suggest some diagrams within the Introduction, e.g., one showing the tasks during flight with the ones you’re focused on in a different color, and maybe a diagram illustrating your four metrics somehow.

Excellent related work section!

Nice annotated event log.

Good clean tables of metrics counts. I’d recommend right-justifying table columns that contain numbers in the future, and keeping numbers of the same type having the same number of decimals.

In this sentence: “This is likely due to language preferences as Flight Team 1 used formalities in only one out of six interactions,” I wondered if you meant Flight Team 2, since it had far fewer formalities in the table of numbers.

Right before the plots there’s an incomplete sentence: “The landing task was the” (bottom of page 14).

Here’s a big point:   
On the plots, how did you establish what the Good Threshold (orange line) was? Also, shouldn’t it be a zone, since Good is not too high and not too low? (-1)

On the plots, it’d be nice to label the points T1 and T2 or have them be different colors and/or shapes.

I like your idea of applying the metrics to the flight phase of a recording that led to a crash to see if you could predict it. But since it’s not clear how you justified your choice for the Good Threshold, the suspicious reader might conclude that the thresholds were simply chosen to help you make your claim about successful metrics.

The definition of the Future Study at the end is quite well done!

Small things (-1)  
- Punctuation follows a citation. There’s no punctuation before a citation, like this (Rodriguez, 2019).   
- typo: corrilaries 🡪 corollaries   
- Spell out numbers less than 10 (e.g., “four” not “4”) unless they’re part of the name of something, like Chapter 1, Figure 4, Flight Crew 2, etc.   
- Maybe a missing word in the first paragraph under Transcript Analysis? I wasn’t sure what Fx3 was; it looked like its explanation might have gotten blended with the Fx2 description.

***Design Studio Team***

**34.5/35 points**

**Comments on report**

The report is very well-written! It looks as if one of you endured a strict English teacher and it paid off. Nice job.

If this were an academic paper, the paragraphs in the Introduction about what studio is and its social aspect would have citations throughout. That’s ok to omit them here, but just wanted you to know.

Great job in the Related Work relating each reading back to your own goal.

Would be nice to have a citation for Gunning-Fog index. (-0.5)

I wasn’t really clear on how you ordered the metrics in Tables 2 and 3.

I was curious how well otter.ai worked for your auto-transcription.

For the section called “The Use of ‘THIS’,” you’d want to use the word “deictic” or “deixis.” That refers to words that are referencing something else, esp. gestures or spatial words.

Have you shared this report with the design studio instructor? I think they might be interested. Or it may be too sensitive with the instructor analysis.

I really like your analysis of the seating.

Great ideas for Future Work.

Overall excellent report!

Small things:   
- You capitalize “Figure 1” just like “Chapter 1.”   
- A citation doesn’t have punctuation before it; it’s more like this (Rodriguez, 2019).   
- In US English, periods and commas go inside quotation marks (in UK English they don’t), e.g., “Wisdom is golden.”

***Combat Team***

**34/35 points**

**Comments on report**

Text is very well-written. I saw very few things I would correct.

I appreciated the signposting in the last paragraph before Section 1.1.

Four videos were analyzed…I’m curious why not more. I’m not complaining, but curious: were there none available or did it take a really long time to analyze each one? Would be interested in your documenting how long the analysis of the 2 hours of video took, and whether you used BORIS or just did it real time with a video player.

I’m interested in whether any of you have military experience.

Tables 1 and 2 are excellent. Nice work!

I’m interested in your Likert-style approach for C-BM2, scored 1-4. I haven’t done that before, but it could certainly work. Might be worth considering 0-3 instead, depending on what stats you’re doing on it. Interested in how you chose 4 graduations.

You tagged metric C-BM2 as decimal, but I’m not sure how that works unless you start averaging. And if that’s true, then why aren’t most of them decimal?

B-BM1 is interesting, b/c it could be bad on your team to have too much of that (“Can I help? Can I help? Can I help?”) or too little. So it’s one of the ones where it’s not just “more is better.”

Some of the metrics, like A-BM1 and L-BM1, depend on how many new situations/new needs for coordination arise, so it might be hard to compare between scenarios.

L-BM2 might give you trouble, because if it’s low, you don’t know if people are rude or just formal and quiet.

“We spoke to some skilled and active game players through an informal interview process” – Would be nice to know how many, and how you judged that they’re skilled. (-0.5)

So, one scorecard per video? Would there be any reason, if you were going to scale this project up, to have each scorecard be allocated to a certain time segment, e.g., each scorecard is for 15 minutes?

I’m not 100% sure how you go from the scorecards in Figures 3-4 to the metrics in Tables 3 and 4.

In Section 5.1 you talk about the metrics being unreliable b/c of not having cameras on many of the instances of communication or events in the scenarios. That’s true, but that’s more of a sensor problem (need more cameras and mics) than a metrics problem, possibly. If you had perfect information, would your metrics be good?

The ICC is more of a measure of your raters’ consistency than anything about your metrics themselves, so I wouldn’t call ICC an assessment of the “reliability of the team metrics scorecard.” (-0.5)

I like your point that EBAT isn’t good at cognitive states.

Need a source for Table 8. I see it’s the Salas Table 21.2 from the slide deck, so I won’t take off points, but you need to cite it.

Excellent analysis!