THE CRUELEST EXPERIMENT

Grandon Gill, Information Systems & Decision Sciences Department
University of South Florida, CIS1040, Tampa FL 33620
ggill@coba.use.edu, 813-974-6755

ABSTRACT

Every year, most of us conduct one or more experiments that employ live human subjects with a potential for major impact on their psychological state and future employability. These experiments are designed without the slightest attention to the safeguards that are normally in place to protect subjects for even the most trivial of research projects. We call such experiments "teaching classes". In a dream sequence, this paper hypothesizes what would happen if a typical course design were subject to the same institutional review board scrutiny as a research project.

Keywords: course design, pedagogy, teaching, computer programming, ethics

PREAMBLE

Not so many years ago, I had a nightmare. Or, if I didn't, I should have. In the dream, I was called before my university's institutional review board (IRB) and asked to justify an experiment I had proposed. I begin this paper by describing that hellish encounter.

THE NIGHTMARE

The nightmare opened with me, sitting in a seat designed for elementary school students, facing a three member IRB panel, resplendent in academic robes. To preserve the sense of anonymity, so critical in achieving rigorous academic review, I will refer to the members of the board as Drs. Torquemada, Sixtus and Carafa, although these were not their real names. Dr. Torquemada led the panel. Dr. Sixtus appeared to be sleeping. The proceedings were as follows:

Dr. Torquemada: Professor, we have called you in to testify before us as a result of a most peculiar experiment that you have proposed. Before passing judgment, we wanted you to speak on your own behalf.

Me: Thank you, members of the panel. *[I find myself squirming uncomfortably in my miniature chair. Who would not be fearful when facing the wrath of the IRB?]*

Dr. T: Before we begin, I'd like to get the basic facts straight. The experiment you propose will last approximately four months and will use student subjects. Presumably, they all volunteered to participate.

Me: Not exactly volunteered. Participation in the experiment is required if they are to graduate.

[*At this point, Dr. Sixtus emits what sounds like a loud raspberry, although it is still not clear that he is awake.]*

Dr. Carafa: We'll return to that later. For now, I'd like to see how you justify your experiment in light of the guidelines promulgated in CFR Title 45, Part 46, dealing with the protection of human subjects. As you doubtless know, the document stresses sound research design, informed consent, and minimizing risks to experimental subjects. Let's start with the research design, shall we?

Me: Certainly.

Dr. T: I'm a bit fuzzy regarding what your dependent variable is. Could you explain?

Me: Of course. The goal of this experiment is to increase each student's knowledge of effective programming techniques and practices.

Dr. **C**: So you're looking at a change variable?

Me: Yes ma'm.

Dr. C: Then perhaps you could explain how your research design measures this variable. Maybe you could sketch it out on the white board.

[*I extricate myself from the chair, step up to the white board and draw Figure 1, as shown.]*

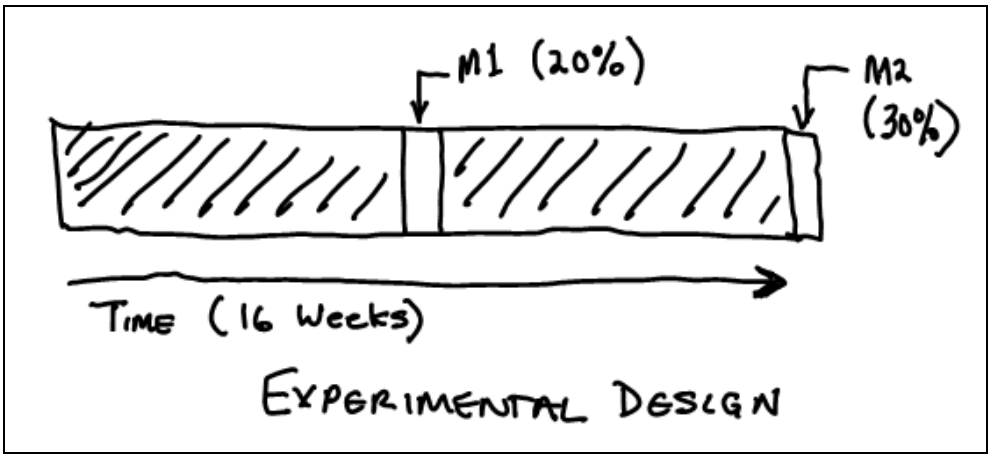

Figure 1: Experimental Design (IRB whiteboard)

Dr. C: Looking at your diagram, professor, I'm a bit confused. You propose that change in knowledge is your dependent variable, while the position of M1 on your diagram suggests that you don't take a measurement until half way through the experiment. And what do those percentages mean?

Me: The measurements use instruments that gauge student knowledge of programming—

Dr. T: Of course! And your dependent variable is the difference between the two measures, reflecting change in knowledge?

Me: Not quite. I don't use a difference. Actually, I add the measurements together using the weights. This way I create a single dependent variable measurement.

Dr. T: Now I'm forced to agree with my esteemed colleague. If the objective of your experimental manipulation is to achieve a change in subject knowledge level, why are you measuring absolute knowledge level?

Me: We make the assumption that subject knowledge levels start at a uniform value of 0. That allows the later measurements to be treated as change variables.

Dr. **C**: I see. And have you attempted to verify that assumption?

Me: Once or twice, but it's harder than it looks.

Dr. **C**: And why is that?

Me: Each time I have tried, the experimental subjects appeared to resist providing a valid measure. My own analysis suggests that they felt they can change the experimental protocol—making it less demanding—by filling out my instrument in a manner that implies they cannot tell a computer from a turnip.

Dr. **C**: Well, if you aren't able to validate your assumption, what kind of support do you find in the literature?

Me: The support is not compelling.

[*At this point, a drop of sweat falls onto my glasses, causing the participants to become distorted and even more threatening in appearance*.]

Dr. **T:** For instance?

Me: Well, Dr. Roberts at Stanford [1] describes differences in backgrounds to be a huge problem, perhaps the major challenge facing computer educators.

Dr. **C**: Then I would have to agree. That evidence provides somewhat less than compelling support for your assumption that backgrounds are uniform.

[Dr. Sixtus, still apparently asleep, emits another raspberry.]

Dr. T: We're starting get off track here. Let's talk about the instruments themselves. What did you call them?

Me: Exams.

Dr. **T**: Exams, then. How have you validated them?

Me: Pardon?

Dr. **C**: Honestly, you call yourself a researcher? All right, let's take it one step at a time. What do your instruments look like?

Me: I'm very proud of them—not a multiple choice question in the bunch. Every instrument I create consists of a selection of short answer, code and essay questions.

Dr. C: And what reliability tests have you built into them—what's the Cronbach's Alpha score, for example?

Me: Truthfully, I've never tried to measure it. You can't really repeat or rephrase questions in the instrument; there just isn't enough time for subjects to complete it if you do.

Dr. **T**: I'll grant you that. But in your own experience, do you find that subjects perform consistently across similar types of questions?

Me: A lot of the time they do. But in scoring the instruments, I have found considerable variation in the ability to answer similar questions by an individual subject. Sometimes the answer comes to them, sometimes it doesn't.

Dr. **T**: But you feel confident that every time you encounter the same subject response, you award the same score.

Me: Not completely confident, there's probably some variation. But I try to be consistent. Also, you've got to understand, these instruments are handwritten so responses can be quite difficult to interpret.

Dr. **C**: Handwritten? But wouldn't that make it possible to identify the subject who submitted a particular response?

Me: I can't imagine bothering to use handwriting do that.

Dr. **C**: And why not, professor high-and-mighty? Are you claiming to be above such petty emotions as curiosity?

Me: Not really. It's just that the subjects' names are on the instrument.

Dr. **C**: They're what?!!!

Me: They have to be, if I am to provide feedback to the subjects at the end of the experiment.

Dr. C: And you're telling me that knowing the subject's identity has absolutely no impact on scoring? Can you think of any instances where it might?

Me: The only concrete example that I can think of is where an answer looks like complete gibberish. If I know the subject to be a strong student, I'll sometimes take a second glance to see if there's any wheat amongst the chaff.

Dr. C: And you'll guarantee to this panel that this is the only source of measurement bias found in your experiment?

Me: I really try to be as objective as possible.
[Dr. Sixtus, still apparently asleep, sneezes with a sound suspiciously like "halo effect" and begins to snore]
Dr. **C**: I'm becoming very uncomfortable with what I'm hearing. And what about inter-rater reliability issues?
Me: Teaching assistants do a lot of instrument scoring—
Dr. **C**: And…
Me: I try to eyeball the instruments for consistency. Naturally, I can't rescore every item. But I think it averages out as being fair in the long run. Also, if a subject comes to me with a question about the scoring procedure, I take a close look at it.
Dr. **T**: Let's keep this process moving. Obviously, we'll have to further address the issue of instrument reliability before we approve this project. But now, I'd like to address the validity of your measurement.
Dr. C: Before we begin, why don't we try this? *[She pulls out a sparkling golden lariat and begins to twirl it. She then tosses it around me and suddenly I find myself compelled to speak the truth. Most alarming…]*
Dr. T: So, professor, how do you address the validity concerns related to your instrument?
Me: If you don't mind, I'll start with content validity. [*My voice becomes stronger and much more self-assured—this is one of my strengths*]. As you doubtless know, Kerlinger [2] [*a choir of angels sings a harmonious chord as the venerated name is spoken*] tells us that content validity tends to involve a considerable amount of judgment. Because I have programmed commercially for years, I am in an excellent position to define survey items that are representative of the universe of programming problems.
Dr. T: Now we're getting somewhere! So completing the items on your instrument will be valid a measure of a student's effectiveness as a programmer.
Me: *[As I try to say "undoubtedly" the lariat tightens around me and chokes my words]* Uh…
Dr. **T**: I didn't quite get that. Let me rephrase the question. So, in your professional opinion, as an expert in the field, you judge your instrument to be representative of the programming process your subjects will encounter in the field?
Me: There are some differences... *[The words come out of my mouth totally without volition.]*
Dr. T: Explain?...
Me: Well, for one thing, nearly all commercial programmers work in groups. The instruments, of course, are completed alone.
Dr. **T**: Anything else?
Me: In a commercial environment, code reuse is emphasized—and the programmer who finds a stash of legitimate, high quality open source code that saves time is a hero. Naturally, we can't have subjects looking for code to copy as they fill out the instruments.
Dr. **T**: I see your point. But aren't the skills that you are testing for with your instruments going to be the same skills that industry is looking for?
[*Strangely, confessing is starting to feel really good. Whether this is a result of the lariat loosening, or some self-destructive urge buried deep in my psyche, I am unable to tell. Suddenly, I can't hold back as I relate my experiences*].
Me: Funny thing about that. Back in the mid-1990s, my department at another institution performed a series of focus groups with senior IS managers from over a dozen firms—all of which were potential employers. When we asked about the skills they considered important, they almost unanimously replied communications skills—not data communications skills, by

the way, just the ability to communicate with others. When pressed for more skills—prompted by us with such statements "wouldn't it be great if they were trained in C++ and could program using the Microsoft Foundation Classes"—they conceded that teamwork skills were also important. When asked about the importance of SQL, they responded that the ability of employees to learn new things on the job was critical as well. In fact, the only employers who seemed really interested in specific technical skills were those who had purchased a software application (usually one that was orphaned, so few commercial training classes were available) and their interest was in having us teach that specific tool.

Dr. **C**: Would you mind summarizing?

[Still standing at the whiteboard, I draw Figure 2, as shown, separating programming specific and general skills with a dotted line]

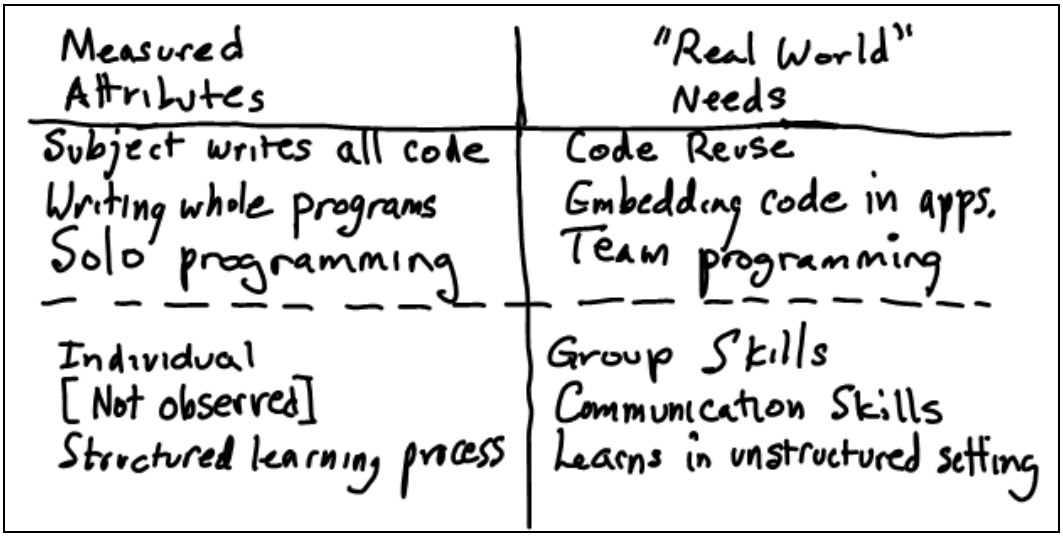

Figure 2: Course Design vs. Practice (IRB whiteboard)

Dr. T: I'm thinking that your instrument and design might need a little work.

Me: (Gulp…)

Dr. T: And you don't have any other measurement tools in your design?

Me: I do, actually. Subjects are required to engage in a number of protocols that I refer to as "projects" and then provide me with copies of their work.

Dr. T: And what impact do these "projects", as you call them, have on your dependent variable measurements?

Me: Minimal, I'm afraid. [*I feel the lasso tightening again.*] Each time I have run similar experiments in the past, subjects spontaneously developed extremely effective interpersonal communications techniques that result in instantaneous exchange of project outcomes. As a consequence, the variability in the dependent variable resulted almost entirely from the instruments we were discussing earlier—project scores were nearly uniform.

Dr. T: So, it seems, we are back to square one. [*There is a long pause, and I find myself sinking knee deep into the floor, which has suddenly become marsh-like in consistency. The only sound is Dr. Sixtus snoring.*]

Dr. C: Moving right along, let's look at the issue of informed consent. The document you gave us—the one you labeled "syllabus"—seems to cover a lot of the issues referenced in 45 CFR Part 46, including duration, explanation of procedures and the sequence of procedures. It

seems to be a bit weaker on the subject of risks and benefits, however. Could you elaborate a bit more on these?

Me: The potential benefits of a degree in my field are great. Currently, of the job categories requiring a bachelor's degree expected to grow over the coming decade, the top 7 are in my field [3].

Dr. T: And you've rigorously verified that completing your course will help students qualify for those jobs?

Me: No, not rigorously.

Dr. **C**: Then would it be safe to describe "the uncertainty that your course will actually help your subjects get a job" as a risk? If so, should they be informed of that risk?

Me: Er, possibly.

Dr. **T**: I think you're being unfair, Elaine. Why don't we take it on faith that taking the good professor's course does qualify students for these great jobs—

Dr. **C**: That would take a great deal of faith, Earl—

Dr. **T**: Enough, Elaine. We'll make the assumption, which takes us to the final key aspect of the IRB investigation, minimizing risks to subjects. What have you done along those lines, professor?

Me: I'm not sure I understand what you're asking.

Dr. **C**: Allow me… What is the percentage of students who complete your experiment at a knowledge level consistent with your minimum expectations, given the manipulations you performed?

Me: Are you asking for a pass rate?

Dr. **C**: Didn't I just say that?

Me: It's about 60%, including withdrawls.

Dr. **C**: And what have you done to minimize failures and withdrawls?

Me: Hmm. We encourage subjects to try their hardest, and give them all the support our severely constrained resources allow…

Dr. **C**: And…

Me: Not much more, really. You see, my experiment is conducted very early in our major. We believe that it makes sense that this experiment should be particularly challenging so that students don't get too far into the major before finding that it doesn't suit them. In a sense, my experiment serves as a kind of gatekeeper.

Dr. **C**: And I assume that you've done rigorous testing to ensure that performance in your experiment is reflective of performance in other offerings in your department?

Me: It depends what you mean by rigorous.

Dr. **T**: Suppose we mean "any"?

Me: Then my answer would have to be no.

Dr. **T**: Okay. I think I've heard enough. The experiment you want us to approve attempts to demonstrate a change in a dependent variable without measuring its initial value. The instruments you are using are of untested—and most likely dubious—reliability with validity characteristics that appear to be diametrically opposed to those of the construct you claim to be measuring. Your attempt at informed consent fails to include a disclaimer that you have no evidence, outside of the anecdotal, that the expected benefits of your manipulation will actually lead to the desirable outcome that you are forecasting in order to induce subject participation. And, rather than actively attempting to minimize risk to subjects, you seem to

be moving in the opposite direction, claiming some form of gatekeeper responsibility. Is this a fair summary?

Me: Ah…It's not quite how I would have put it, but—

Dr. **T**: Then I have just one more question to put to you: what department had the nerve to send you to us with this proposal?

Me: MIS, sir…

[*A deafening raspberry is emitted by the still sleeping Dr. Sixtus*]

Dr. T: MIS? Why didn't you say so? [*He suddenly sprouts a huge white beard and his robes transform into a red suit, trimmed in white fur with a wide black belt across his ample midsection*]. Ho! Ho! Ho! If we didn't cut you folks huge amounts of slack, there'd be no published MIS research at all! Go ahead, professor, conduct your experiment.

[*A samba band now appears in the room and Drs. Torquemada and Carafa begin to dance the lambada. Dr. Sixtus awakens and morphs into a giant capybara, scrambling around the room squealing "Research rules!" Subsequent events in the nightmare, being relevant only to my mental health professional, are omitted.* ***Postscript:*** *Also omitted is the entire analytical section and conclusions of the original paper, which featured my (presumably) doomed attempts to justify the myriad of academic submission norms that I violated in the first half of the original paper (which I present, unaltered, for your astonishment). There were a number of reasons for choosing to write the proceedings version in this manner. First, I was not exactly sure how one goes about writing the "Cliffe" notes version of a nightmare sequence—particularly since dropping purposeless devices such as IRB-panelist-to-rodent transformations would have severely detracted from the essential dignity of the piece. Second, one of the original reviewers—bless his or her heart—suggested that the nightmare sequence could stand alone. Being of a particularly timid disposition, I had not had the courage to do this originally—but when I read the review I found myself wishing that I had. Third, including the full original paper would have represented a serious intrusion on the page numbering space of submitter 415. Finally, I did not wish my colleagues to think me so vain (or so desperate to see my own words in print) as to have spent $450 in excess page charges to put my full work on CD. I'd be enormously grateful, however, to anyone who expresses the slightest interest in perusing the original, and would gladly email them a copy post-haste*].

REFERENCES

[1] Roberts, E. 2000. "Strategies for Encouraging Individual Achievement in Introductory Computer Science Courses". *SIGCSE 2000, 03/00 Austin TX*. 295-299.

[2] Kerlinger, F.N. 1986. Foundations of Behavioral Research, 3rd Edition. New York: Holt, Rinehart and Winston.

[3] Bureau of Labor Statistics. 2004. Occupational Outlook Handbook, 2004-2005 Edition (Table 1), accessed at http://stats.bls.gov/news.release/ooh.t01.htm on 30 March 2004.

ACKNOWLEDGEMENTS

I was quite astonished when my original submission was not summarily dismissed by the reviewers, and flabbergasted when it received the "Distinguished Paper" award for the DSI Innovative Education track. All I can say to the reviewers is that if your objective was to energize me, get my creative juices flowing and give me a new respect for my colleagues, you succeeded admirably!