Keeping Your Student Ratings of Instruction Behavioral with Frequency and Celeration Analysis

Stephen A. Graf

Youngstown State University

First of all, just for the record, I'd like to know how many of you feel familiar with the **standard celeration chart**? Raise your hand please. And how many are not familiar with the **standard celeration chart**? Okay. And, starting to refine this question a little bit, how many are fluent with the either with both the Standard Celeration Chart Flashcards and the Learning Picture Flashcards? Okay.

To sort of preface this, the ratings that I am talking about are the
types of student ratings that seem to be a widespread phenomenon in
colleges and universities across the country. To get a little picture of how 1m
we stand with respect to our various backgrounds or where we are right
now, how many are in a position at this point, at least not this week maybe,
but this year, where you are being rated by some sort of student
evaluation system? Okay. And how many of you are in a position where
you are actually making ratings of instructors or some other positions?
Okay.

Well, the point that I'd really like to sort of start and end with is, 2m
and my essential point is, that *all of this is essentially irrelevant*.
However, nobody's doing the relevant stuff, and they're also doing the
irrelevant stuff wrong. So as a first step trying to give you some ideas on
how we might do a little better job of working with the irrelevancies, then

later get on with more relevant issues. That relevant issue, hopefully somewhat simply stated is this: That what we're really interested in I believe, and I think most behavior analysts agree, is in the teaching situation, what performance the learner comes away with after the instruction. That's a very different measurement than how the learner 3m
feels about the performance of the instructor. So the actual skill or performance or the learning of the learner is not what's being measured by these student ratings. And I think it's absolutely essential to make that very, very clear at the outset. So these ratings are a good deal different than actual student performance.

When we look at how we assess instruction this is perhaps a useful component. But when used as the sole measuring device of how the teacher is doing in the classroom, then I think it falls extremely short of anything relevant whatsoever. Okay.

The first thing that I'd like to show you is a set of data for three 4m
instructors. I obviously wanted to retain anonymity, so I did away with the names when I collected the data. I then put on fictitious names, but in re-thinking that I've done away with those also because I realized there are biases against female names, biases against ethnic names, and some of my fictitious names fell into both of those categories. And there may be other biases as well. So, here is the data that I'd like to present to you. What we're going to do is sort of put ourselves into a situation where we 5m
need to make some sort of a decision based on whatever measurement we have of whether these three individuals are going to get whatever: Promoted, probably a useful sort of thing to consider for this, but other types of decisions would involve tenure, reappointment , and heaven forbid, probably within a few years, retrenchment (where you decide who

you are going to keep and other people get let go, not just not promoted).
Well, here is the data that I'd like for you to make your decision from. (LL)
Now, that represents data from three people. It all sort of should be on the
same line, but they're so close together that we couldn't quite do that very 6m
effectively. The time that we're talking about here is, notice it's in 1980
which is, so this is that particular time was appropriate collection of this
particular data. Now, the question is, how are you going to make your
decision? That is, what are you going to use? What we're doing is sort of
hypothesizing or assuming that these individuals on things such as
scholarship, research, university service, some of the other indicators that
are typically used in these decisions are essentially equal across all three
individuals. So, it's boiling down to something like this. And now how are
you going to make the decision?

Well, for our uses we're calling these people 1, 2, and 3 from left to 7m
right. So, if you have any ideas as to which one is best, based on what you
can see there, then you should write that down. In fact, you should write
down, or at least write in the air, something. Either 'no decision', or 1, 2, or
3. Okay, so please do that now. And then sort of share it with your
neighbors as to what decision you came up with. (PAUSE)

Now, the next question I should address is how were these particular
data produced? Well, I think they're rather typical, because in effect what
is involved here is some sort of a questionnaire, where five different
possibilities have been presented, and the individuals that are being rated 8m
are supposedly the subject of the rating. The students do the ratings and
they're responding to some general question. And their responses, then,
are to be put into any of those five categories. Or they can opt just to
forget about it and not make any response at all. And alot of the students

take that as an option. So, the top one might be something like "outstanding." Number 4 might be "very good." Number 3 might be "good." 2 might be "adequate." And 1 might be "inadequate." And these in fact were the wordings used on the ratings that were generated for these three dots.

Now, the question is, of course, why just one quarter? Obviously you 9m
don't want to assess individuals simply on one little spot in time. That seems to be fairly ludicrous. But, as a matter of fact, what is happening is these three dots represent mean of means. That is, these people get a rating once a year, at least once a year, sometimes more often, and they get a mean where these are the weights. And the number of the responses in the 5 weight are multiplied by 5, and the number of responses in the 4 weight are multiplied by 4. These are added up and divided by the total number of responses. I think most of us are familiar with that sort of a situation. What I'd like to point out to you is that that really is
inappropriate measurement. First of all, the use of 5,4,3,2,1 is really a 10m
scale which is not nominal, and yet is, these types of categorizations are very clearly nominal categorizations. So that, while they're equal interval, they're not even equal ratio. And if we assume equal interval scales then we're probably making an inappropriate assumption as well. The distance, for example, from 1 to 2 is not necessarily the same as the distance between 2 to 3; it's not necessarily the same as the distance between 3 to 4; it's not the same as from 4 to 5. So get in very great danger with that. So it really makes no sense at all to weight these types of categories.

The other point then, of course, is that once you have a mean for a 11m
particular quarter, then, throw out all of your time data by squishing all of that data across time together into one mean of means is also a rather

presumptuous and really ridiculous practice. And yet it is very commonplace. And for the individuals who are represented here, that was in fact exactly what was being done as the current practice within that particular department of the university. And so those individuals had to live with it.

The next thing I'd like to do is start giving you some of the data that
I collected on myself prior to looking at some other individuals. With this
type of information, self-recording is easily obtainable, and yet it's not that 12m
often done, many instances simply because the people don't appear to look
at the possibilities that are available. That's one of the other things that I
think would be an appropriate point to make here today. And that is, that
you have some options which you might do to make your own ratings a
little bit more behavioral by using some of the techniques available. And
this would be one.

I've put in these phase change lines to indicate that the 5 through 1
category names or labels often change. And such a change occurred in
1977, between the Summer and Fall quarter of that year. So that, whereas 13m
previously 5 was "excellent," 4 was "good," 3 was "average," 2 was "poor,"
and 1 was "very poor," it was then changed so that 5 was "outstanding," 4
was "very good," 3 was "good," and etc. The interesting thing was that the
overall university mean prior to this phase change was 4.0, on this 5 scale.
And the overall, as would be anticipated, the overall mean afterwards was
also 4. So, changing the categories does not seem to make much difference
in the way the scores come out. Now I have heard of a couple of instances
where you can set up the categories so that that will occurr. And one is the 14m
best I ever had at the top of the scale. That will sort of change things, shift
things a little differently.

Well, in closing then, the means from my own data. These are mean
student ratings every quarter, across a period of approximately five years.
And again, the ceiling is 5. You can't get anything better than a 5. And of
course the "poors" is 1, can't get any worse than 1. Fortunately I never got
too close to "poor"; starts out pretty low. Then these are the median rating 15m
across that same time. Once again these are using that 1 through 5 scale
and getting the middle, either by mean or by median. Well,...

ORL: Was it a university wide median?

SAG: The university wide median was 4.

These are my own medians per quarter.

The other thing which is typically not done, and that's collecting the
data every quarter. I really don't feel comfortable in connecting these data
points because there's a lot of no-chance opportunities in there where the
students aren't making these types of reactions. And yet, for the way that
the university policy is set up, the full time professor has to undergo these
only once a year. And that is in fact what in many cases exactly what they 16m
do is have ratings done once a year. So, for 10 data points it would take 10
years. At 3 a year you get there a little more quickly. Once again, these
are very non-behavioral. I've indicated that by the question marks down
at the bottom, because these are not students making ratings. These are
abstractions, completely figments, and are derived measures, which are
really not based in behavior at all.

So the next thing is to sort of take a look at how we might go about
making these more behavioral. One thing to do is to simply convert the
numbers in the categories to counts per quarter. And look at them 17m
individually. We can start by simply looking at the number of students

who are making ratings. This is a count across time. These are actual numbers. These are students actually behaving. That is, now we have students as the behaviors, and what they're doing are making ratings. And I've drawn now a celeration line through the data for these particular points. These were students making ratings. As many of you know who are involved with these ratings, there also are students who do not make ratings, students who've been around a bit, and students who do something like skip class on the day or the time that the ratings are given. And so 18m
this is a look at students not making ratings. Students doing something else other than rating. (PAUSE)

The next point I'd like to bring up is that with respect to this type of a picture it became fairly clear in the first several quarters that the students themselves were becoming very disenchanted with the evaluation process. And I found that it was necessary to sort of arrange the contingencies of the course, such that very rewarding things were 19m
happening at the same time, or on the same day at least, that the student evaluation, so-called student reaction to instruction were taking place. So we had the 'Graf All-Star' awards for student performance in the course, being given prior to the rating. Still lost some students because once it happened why they took off. But . . .

ORL: What was it?

SAG: What were the?

ORL: What was the 'Graf All-Star Award'?

SAG: It's a little certificate, with a miniature **standard behavior chart,** shrunk down, which said that the student had earned an A in the course prior to the end of the quarter, and didn't have to take the final, and it was hoped that such appropriate actions would continue in their general life as well as other courses.

A: Did they still have the course final (?) ?

SAG: They came to class to find out. Their results of their exit assessment 20m
were also given on that particular day too, and several things that were happening. So, they were sort of in a state of not knowing before they came to class of how they'd done on the exit assessment, whether or not they'd have to take the final, and whether they made the 'Graf All-Stars' or not.

Well, what I'd like to run through now are the actual one through five plots for the various categories. This again started out as the "very poor" category; this is category 1. Students responded "very poor" to the
instructor's performance. And as you can see this is a /1.4 every six 21m
months.

So, these again are simply slicing off one of the categories in that five category tier of student reaction to instruction.

Now it's appropriate to keep as an underlay, the number of students who are actually making the ratings, because that's sort of our baseline celeration. So, the number of students making ratings is going up while this is going down.

Next category (TR) 2, which was originally "poor," and turned into
"adequate." (P) And then category (TR) 3. And category (TR) 4. And finally 22m
category (TR) 5. (P)

When we take a composite look at the celeration lines for each of the 5 categories it seems that the general statement that one could make is
that each of these categories is really independent. That is, when we're 23m

taking a mean we are falling into the trap of losing alot of information about the data. And the way we really should look at it if we're going to have to deal with category scales such as this is to actually keep each of them separate. A separate chart for each one.

As sort of an incidental, I think it could be argued from the fact that
the high ratings were going up, and were accelerating most steeply of any
of the categories, that grade inflation may have been involved, and Graf
was just turning soft and giving more and more A's. Well, one way to 24m
↑ explore that is to look at the frequencies across quarters of the number of
grades that students earned in each category. First, the total number of
students who earned A through F. Then, the number of student who
earned F's,(TR) then the number of students who earned D's,(TR) C's,(TR) B's,(TR) and finally
A's.(TR) ↑ It is true that the A's did accelerate slightly. There's a rival 25m
hypothesis and that is the instructor became more proficient in helping
(LL) the students achieve their target behaviors.

A: Did they find out their grades before or after the ratings?

SAG: They found out their grades before they rated, or at least their grades prior to the final. So, some individuals had clinched A's. If they had not clinched an A they could have clinched at least a B, but they still had an opportunity to get an A, and so forth down the scale.

ORL: Even if they were related that doesn't imply they were causal.

SAG: That's correct.

ORL: In a real good course you receive high grades of students () and high grades high ratings from professors.

SAG: Yeah.

ORL: They're real important.(?)

SAG: Yeah. I guess what I'm saying is that there does not appear to be any

rampant grade inflation going on here. 26m
ORL: If they're learning more and more then it should have been.
SAG: Yea, that's right. They're not. Yeah, I think that's very, very
definately the catch in this particular course the students are not getting,
the students across quarters is not performing a great deal better than
prior students. A little bit of an increase, but not all that much. So, you
kind of win on grade inflation but you lose out on performance that counts,
if the grades are based on student performance.

Well, the thing that seems to me to be a very likely possibility is that 27m
this type of a picture is not going to be very satisfactory to too many
people. There are really too many things going on and I think that's really
the fault of the instrument. That is, we know we can get a decent two line
learning picture , and three is still giving us some difficulty trying to deal
with it. But how would that look if what we did was now just take the
number of students making ratings each quarter, and then have two
categories: "Outstanding"s if you wanted to sort of select off that top
category, and then lump everything else as simply "not-outstanding." And 28m
that's what I did next. And that's really I think what the easiest course
would be for the next step. So. (TRS) (PAUSE 28 sec.) If we take out for the
moment the actual number of students making ratings then this is what 29m
we're left with on the "outstanding" versus "not-outstanding." ((And I'm
glad Ray Beck isn't here because I did this before I found out about the red
pen. And I apologize if anybody is colorblind for using red).) Okay, well,
the final look then is sort of a composite of that particular data and the
green line is the celeration of students rating the instuctor "outstanding."
The purple line is students making some rating other than "outstanding."
And the black line is the total number of students making ratings.

So, if we then go back to our original situation where we had three 30m
instructors who came out looking like this on the usual type of
measurement system proposed. If we now look at each of those three with
'rating pictures' then this is what it looks like. First of all, this is the
number of students who made ratings for Instructor number (TR) 1. And the
green line again is the number of students who rated the instructor
"outstanding." The purple line, number of students who rated the
instructor something else other than outstanding. That is Instructor 31m
number (TR) 1. Instructor number 2: (TR) (P) Students make ratings.
Students--green line is students who rated the instructor "outstanding,"
purple line students rate instructor "not outstanding." (P) (TR) And finally for
Instructor number 3, or professor number 3, (TR) (P) number of students 32m
making ratings, celeration of that, and the "outstandings", and "other than"
ratings for Instructor number 3. (P)

So, now what I'd like to do is to go back for a moment and take away
our learning pictures of the three instructors, and have you have a second
opportunity to make your decision if you are going to make any decision at
all. This now is a re-hash of the celerations for instructors number 1, 2,
and 3. And the black line is students making ratings, the green line is 33m
students rating the instructor "outstanding," the purple line students rating
instructor something "other than outstanding." I'd like for you to take a
pencil, or write in the air, whisper to your neighbor, I want you to make up
your decision which of these instructors now you are going to promote. (P)
And you may also want to give some rationale to share that with your
neighbor. And I'd like to see how it comes out. (P 5sec) Who's still not
had an opportunity to make your decision? Okay. By show of hands, how
many chose Instructor number 1 to get the promotion? How many chose 34m

Instructor number 2? And how many chose Instructor number 3? Okay.

What I'd like to ask at this point, is who has a rationale for the decision that they made and are willing to share with us? Carl?

Carl: The improvement index. the relationship between the black line and the green line, the total number of people, the celeration of the number of "excellents."

SAG: Okay, the improvement index?

Carl: Or the improving proportion.

SAG: Okay, so you got. And which instructor did you choose on that basis?

Carl: Number 3. 35m

A: Number 2 which is the one that you might choose Steve: The number of "outstandings" is growing more or less with the number of students in the course.

SAG: Who had the same sort of rationale? Who had a different rationale? And for those who did not choose number 3, what rationale did you use for your selection, if you're willing to share it with us. (P) Okay.

Well, that's the ideal situation that I am proposing. Now what I'd like to do is to kind of go back to the old world, to the real world, and what I've done is scramble those professors, 1, 2, and 3. And now we're going to call them W, X, and Y. And they're not necessarily in the same order. And since many people on promotions committees don't like to look at charts, 36m
and don't always just look at the means, which is what we looked at first, now I'd like to give you an opportunity to make a decision based on the actual frequencies of professor W, professor X, or professor Y. And I'd also like for you to raise your hand when you have finished your decision.

(LL) (PAUSE) 37m

SAG: Who is finished? Who's given up? (LL)

Bea Barrett?

A: How can you be finished, Steve?

SAG: How many are going to choose professor W? How many are going to choose professor X? How many are going to choose professor Y? Okay. What I submit is that the confusion which we expressed here is exactly what's going on in our college and university promotions committees, and I think for at least this reason if not some others. This W was 2, X was 3, 38m
and and Y was 1.

That concludes my presentation. Thank you very much. 38m 9s.

NOTE: USE THE FOLLOWING REFERENCE WHEN REFERRING TO THE ABOVE PAPER:

Graf, S.A. (1981). Keeping your student ratings of instruction behavioral with frequency and celeration analysis. Invited Address presented at the meeting of the Association for Behavior Analysis, Milwaukee.

Presented Saturday, May 30, 1981 11:00-11:50 a.m. (actual running time was 38' 9").

Sweeney, C. Chaired the Invited Address.

(Seventh Annual ABA Convention, May 28-31, 1981)

Transcription Conventions used in the above document:

() - Unintelligible audio -- text could not be discerned.

(?) - Best guess at spoken wording.

(LL) - Audible Laughter by audience members.

(PAUSE) or (P) -- Noticeable pause in talk. No speaking for duration.

(TR) - Transparency audibly placed on overhead.

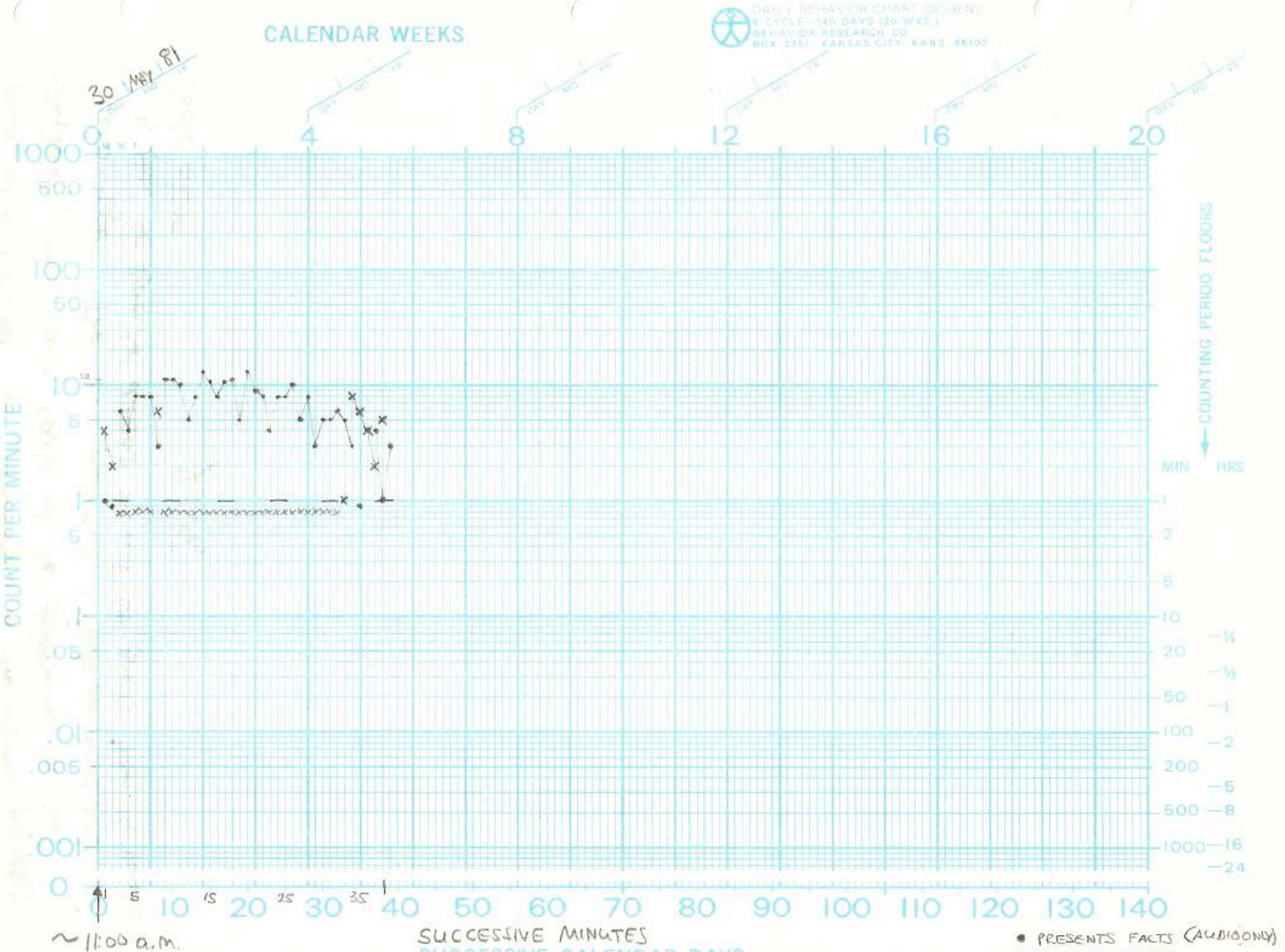
SAG: - Stephen A.Graf (responds to questions).

ORL: - Ogden R. Lindsley, in audience (asks questions, makes comments).

Carl: (Carl Binder??) (makes comments).

A: Unidentified persons in Audience (asks questions, makes comments).

↑ 1.00 - Mark and indication of each successive minute in the Talk.



~ 11:00 a.m.

• PRESENTS FACTS (AUDIO ONLY)
X MANDS AUDIENCE

KEEPING YOUR STUDENT RATINGS OF INSTRUCTION BEHAVIORAL

SUPERVISOR ADVISER MANAGER

STEVE GRAF
BEHAVER

AGE LABEL COUNTED
in Presentation at ABA
May 30 1981

SEE BACK
DEPOSITOR AGENCY

TIMER

COUNTER

J.W.E.
CHARTER
6-11-86

NOTE: The frequencies of facts per minute presented pertains to spoken, audible facts only. Many visual facts were presented via transparencies on the overhead projector. Obviously, this count could not be determined from an AUDIO tape. The true picture of "facts per minute presented" will, then, be somewhat different & will be greater than the audio facts alone.

J WE.