THIS DOCUMENT IS A ROUGH DRAFT. PLEASE EXPECT CHANGES IN FUTURE VERSIONS.

Eric Kernfeld and Paul Sampson

UW Statistical Consulting program, January 2016

Analysis Plans for Transparency Data

This document describes statistical approaches for analyzing results of a human-robot interaction study. This is a technical memo, not a full record; it is intended to be used by those familiar with the project. For more information on the project, contact Leah Perlmutter. If you are a stat/biostat members with access, see the consulting program's winter 2015 records.

Scientific Context

These analysis plans aim to test three hypotheses:

- H1: Adding visualization-based transparency after using only natural transparency will improve task metrics.
- H2: By improving the user's mental model, visualization-based transparency will improve task metrics even after it is removed.
- H3: The medium in which visualizations are provided (monitor versus headset) will not impact task metrics.

They are, however, constrained by the study's design. The document is broken into two parts:

- Analysis 1 uses uncontroversial assumptions, but it can test only H3, not the others.
- Analysis 2 assumes the effect learning is constant over time on a logarithmic scale. This assumption allows for testing H1 and H2.

Implementation

One easy way to implement the analyses described below is to use the `R` computing language and the package `lme4`. This combination of software allows models to be specified in *formula syntax*, which is similar to the descriptions below.

Data setup

These data were gathered on 20 subjects. For each subject, multiple responses were measured:

- number of attempts averaged over a set of tasks
- number of words for each attempt averaged over a set of tasks
- time taken for each attempt

ANOVA-style linear models generally are more reliable in practice when data such as these are log-transformed. These measurements are all strictly greater than zero, so taking logs is no problem. In the description of the analyses, Y_i will denote the log of the task metric. To describe the individual, i will range from 1 to 20. To complicate things slightly, each task metric was measured for baseline, monitor, and headset, and for phase 1 (P1, visualization present) and phase 2 (P2, aftereffects). This will be described as Y_{ij}, so that:

- Y_{21} is the log task metric for person **2**, P1, baseline
- Y_{11} is the log task metric for person **1**, P1, baseline
- Y_{12} is the log task metric for person 1, P2, baseline
- Y_{13} is the log task metric for person 1, P1, monitor, *even if the monitor came after the headset for that person*
- Y_{14} is the log task metric for person 1, P2, monitor
- Y_{15} is the log task metric for person 1, P1, headset, *even if the headset came before the monitor for that person*
- Y_{16} is the log task metric for person 1, P2, headset

For each individual, let X_{ij} be:

- 1 if $j = 3, 4$ and person i used the monitor last
- 1 if $j = 5, 6$ and person i used the headset last
- 0 otherwise

Analysis 1

Model and interpretation

The proposed statistical model is a *random-effects* model. It takes the form

$$Y_{ij} = z_i + \mu_j + \beta_{learn} X_{ij} + \epsilon_{ij}.$$

To interpret each component:

- $\epsilon_{ij} \sim N(0, \sigma^2_{phase})$ -- A random "noise" term describing natural stochasticity in a single measurement Y_{ij}.
- $z_i \sim N(0, \sigma^2_{person})$ -- a random effect specific to person i, describing their initial aptitude with the robot.
- β_{learn} -- the average effect of learning that occurs between the first transparency device and the second (rounds 2 and 3; tasks 16 and 31; batches 7 and 13). More info in the section below.
- μ_j -- a fixed, unknown parameter describing:

- $j = 1$: the average log task metric for P1, baseline
- $j = 2$: the average log task metric for P2, baseline
- $j = 3$: the average log task metric for P1, monitor, *having learned from the 15 baseline tasks*
- $j = 4$: the average log task metric for P2, monitor, *having learned from the 15 baseline tasks*
- $j = 5$: the average log task metric for P1, headset, *having learned from the 15 baseline tasks*
- $j = 6$: the average log task metric for P2, headset, *having learned from the 15 baseline tasks*

Capabilities and Assumptions

This model assumes measurements from individual people are statistically independent. This assumption is crucial.

The model assumes the effect of learning between rounds 2 and 3 is *multiplicative*: on average, the extra practice alters task metrics by a certain percentage. If people learn so much from the headset that their monitor performance skyrockets, but not the other way around, this assumption would be violated.

This assumption brings us to tests of H3. It is possible to estimate every parameter in this model, in particular μ_j. The parameters μ_j cannot separate the transparency's effect from the effect of learning during the 15 baseline tasks; Nevertheless, testing $\mu_3 = \mu_5$ or $\mu_4 = \mu_6$ gives a reliable test of H3. This assumes that the effect of learning during the first 15 tasks is equal (on average) between the monitor-first group and the headset-first group.

As a side note, μ_4 and μ_3 are not directly comparable, because of the differences in batch size between P1 and P2. This is true for any pair of j values where one is odd and the other even.

This paragraph describes a test not outlined in H1, H2, or H3, but the parameter β_{learn} can be tested. If it is a statistically significant, that means the learning effect is probably not due to chance. Since β_{learn} is added to the log task metrics, $e^{\beta_{learn}} = 0.7$ means that learning between rounds 2 and 3 decreases average task metrics by 70%. This can be used to check whether the effect of learning is scientifically meaningful.

This analysis plan does not address H1 or H2.

Analysis 2

The statistical model here is very similar, but it makes one big assumption about the learning curve and it requires one additional covariate. Let X'_{ij} be 1 if $j = 3, 4, 5, 6$ and 0 otherwise. The model is:

$$Y_{ij} = z_i + \nu_j + \beta_{learn} X'_{ij} + \beta_{learn} X_{ij} + \epsilon_{ij}.$$

To interpret the components that differ from Analysis 1:

- β_{learn} -- the average effect of learning that occurs between the first transparency device and the second *or* *the baseline and the first device.*

- ν_j -- a fixed, unknown parameter describing:
  - $j = 1$: the average log task metric for P1, baseline
  - $j = 2$: the average log task metric for P2, baseline
  - $j = 3$: the average log task metric for P1, monitor, *without the effect of learning from the 15 baseline tasks*
  - $j = 4$: the average log task metric for P2, monitor, *without the effect of learning from the 15 baseline tasks*
  - $j = 5$: the average log task metric for P1, headset, *without the effect of learning from the 15 baseline tasks*
  - $j = 6$: the average log task metric for P2, headset, *without the effect of learning from the 15 baseline tasks*

Capabilities and Assumptions

This model makes the same assumptions outlined in Analysis 1:

- It still assumes measurements from individual people are statistically independent. This assumption is crucial.
- It still assumes the effect of learning between rounds 2 and 3 is *multiplicative*.

One extra assumption is that the effect of learning between rounds 2 and 3 equals the effect of learning between rounds 1 and 2. We described this by saying "learning is additive", but since we are on a log scale, this should be reworded. The model assumes *the percent change in average task metrics due to learning between rounds rounds 2 and 3* equals *the percent change in average task metrics due to learning between rounds rounds 1 and 2*.

Since H3 can be tested without this extra assumption, Analysis 2 does not address H3. For the rest, the test $\nu_3 + \nu_5 = 2\nu_1$ can address H1. The test or $\nu_4 + \nu_6 = 2\nu_2$ can address H2. These tests only assess whether results are due to chance. To assess the scientific importance of the effects, exponentials of differences are useful. For example, $e^{\nu_3 - \nu_1} = 0.7$ means the estimated *effect* of the *monitor* is to reduce task metrics by 30% on average. For another example, $e^{\nu_6 - \nu_2} = 0.7$ means the estimated *aftereffect* of the *headset* is to reduce task metrics by 30% on average.