Supplement for Improving Comprehension of Measurements Using Concrete Re-expression Strategies

USER STUDY: MEASUREMENT COMPREHENSION

We report in the paper that we conducted a within-subjects user study on Amazon Mechanical Turk, adapted from the estimation study of Barrio et al. [1], to answer the question, *Does viewing our concrete re-expressions help users estimate new measurements more accurately?* We describe the study implementation and results in more detail.
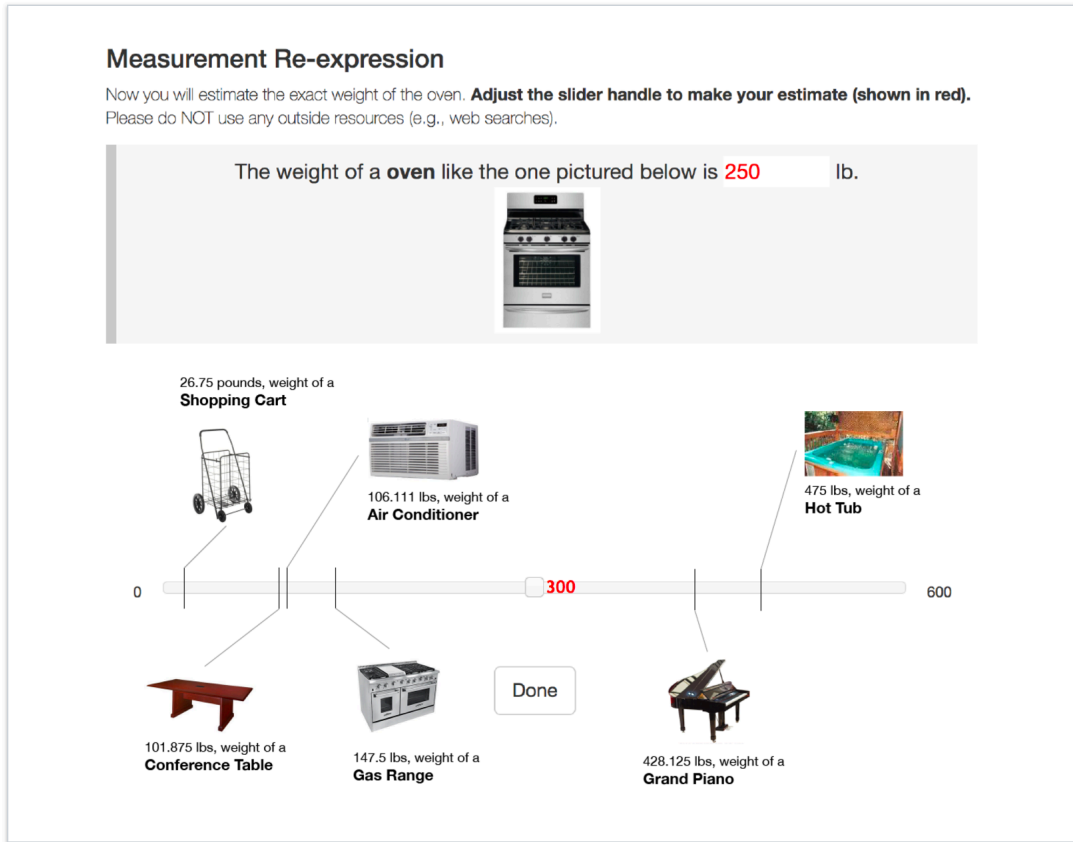

Figure 1: Interface for measurement comprehension study. We present an image of an instance of a synset (i.e., a specific Amazon product) and ask the participant to estimate the object's weight, height, length, or volume. In this example, the participant sees re-expressions from our Adding Familiar Context algorithm as they make a guess (i.e., move the slider).

Stimuli and Procedure

Participants were shown images of objects and asked to guess their weight, height, length, or volume with and without the aid of our concrete re-expressions. We adapt the estimation study used by Barrio et al. (2016) in which participants are asked to provide a missing measurement and get either no re-expressions or the benefit of dynamically generated re-expressions as they enter their guess.

We selected eight specific examples (i.e., Amazon products) of synsets in our database: a chest of drawers, gas range, grand piano, oven, cradle, sawhorse, hand truck, and violin. From these eight examples we created 16

trials, where each trial presents the participants with an image of the object and synset name, and asks them to provide the weight, height, length, or volume by positioning a slider handle, either with or without the benefit of viewing re-expressions as feedback for the guesses they make. Only the slider was visible for entering a guess for No re-expressions trials. For Re-expression trials, we varied whether the re-expressions represented the Adding Familiar Context or Reunitization strategy between subjects.

Trials were blocked into Re-expression versus No re-expression blocks and counterbalanced across participants, such that some participants saw the block of ``None'' treatments first, followed by re-expression treatments, and the same number of other participants saw the blocks in the opposite order. We ensured that a participant never saw any of the objects they were guessing about in the trials in the re-expressions that were generated as they moved the slider to make a guess. To set the range of the slider for each trial, we randomly chose a slider range l that was between 1.25 and 5 times the true object measurement so that the position of the true answer varied. We then randomly set the initial slider position.

We implemented the study as a HIT containing 16 trials on Mechanical Turk with a reward of $2.00 and bonus of $1.00 for quick answers to discourage use of outside sources.

Results
120 workers completed the task, taking 7.0 minutes on average. The mean accuracy (absolute error of estimate) was lowest for Adding Familiar Context (mu: 0.285, sigma: 0.243), followed by Reunitization (mu: 0.347 sigma: 0.282). The mean accuracy for the ``No Re-expression'' condition was 0.35 (sigma: 0.32).

To compare the ratings by treatment taking into account the repeated measures design, we ran a mixed effects linear regression to regress the absolute error term on treatment (Adding Familiar Context, Reunitization, None), including task id and token (unique per worker) as random effects and the order of blocks and order of trials within a block as fixed effects. See the .Rmd and .html files in this directory for model results and confidence intervals for effects. To summarize, we find that viewing an Adding Familiar Context re-expression improved accuracy, but Reunitization did not. We speculate that this lack of effect may result from the mental work it requires to interpret the re-expression objects as a multiple of the slider input, which is changing frequently.

REFERENCES
1. Pablo J Barrio, Daniel G Goldstein, and Jake M Hofman. 2016. Improving the Comprehension of Numbers in the News. *Proc. CHI 2016* (2016).