**1. Study information**

**- title**

Investigating contextual effects on the production of English probability expressions

**- authors**

Michele Herbstritt, Michael Franke

**- research questions**

Under the assumption that a sentence of the form *probably p* is true in a given situation if and only if the likelihood of *p* exceeds a certain threshold theta (Swanson, 2006, Yalcin, 2007, 2010, Lassiter, 2011, Moss, 2015), we investigate three possible sources of variability in English native speakers’ production of such sentences.

1. does the actual likelihood of *p* affect speakers’ choice of probability expressions (e.g., from a fixed set of options: *certainly (not), probably (not)*)? For example, are speakers more or less likely to say *John will probably win the lottery* when *P(John wins)=0.5* than when *P(John wins)=0.3*?
2. assuming a fixed likelihood *P(p)=x,* are speakers’ choices affected by the configuration of the state space in terms of different possible events alternative to *p*? For example: fixing *P(John wins)=0.5*, are speakers more or less likely to say *John will probably win the lottery* in a situation where *John wins* is the most likely event among a set of three or more alternatives (call this 'plural' configuration) than if *John wins* is one of exactly two alternatives (call this 'dual'), even if *P(John wins)=0.5* in both situations?
3. assuming a fixed likelihood *P(p)=x*, are speakers’ choices affected by the kind of question under discussion (QUD) in the conversation (Roberts, 1996)? For example: fixing *P(John wins)=0.5*, are speakers more or less likely to say *John will probably win the lottery* in a situation where they take their statement to be answering a *wh-*question such as ‘Who will win the lottery?’ rather than a *polar* question such as ‘Will John will the lottery?’, even if *P(John wins)=0.5* in both situations?

**- hypotheses**

*(For each of the research questions listed in the previous section, provide one or multiple specific and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here.)*

The research questions are investigated in the context of a multiple-choice task. Speakers’ choices are restricted to sentences of the following four forms: *certainly not p*, *probably not p*, *probably p* and *certainly p*. In general, we expect each of these expressions to be more or less pragmatically appropriate to describe different situations. More specifically, we put forward the following three hypotheses.

1. given the semantic assumption outlined above, we expect the likelihood of the event *p* to have an effect on speakers’ production of probability expressions. For example, all other things being equal, we expect a likelihood of 0 to be associated with comparatively higher choice rates of sentences such as *certainly not p*; likelihoods between 0 and 0.5 to be associated with higher choice rates of *probably not p*; likelihoods between 0.5 and 1 with higher rates of *probably p*;finally, likelihood of 1 with higher rates of *certainly p*. This translates into a hypothesis about our regression model (see below), namely that the coefficient for factor VALUE be credibly different from and in fact higher than 0.
2. given the literature on the so-called Alternative Outcome Effect (Windschitl & Wells, 1998) we expect different configurations of the state space to correspond to different production patterns of probability expressions. In particular, all other things being equal, fixing a likelihood of *p* sufficiently bigger than 0 and smaller than or equal to 0.5, we expect plural configurations to be associated with comparatively higher choice rates of *probably p* and dual configuration to be associated with comparatively higher choice rates of *probably not p.* This translates into a hypothesis about our regression model (see below), namely that the coefficient for (non-reference level) SCENARIO:PLURAL should be credibly different from and if fact higher than 0.
3. building on intuitions about the meaning of *probably* contained in (Yalcin, 2010) and (Lassiter, 2011), we expect different kinds of QUD to correspond to different production patterns of probability expressions. In particular, all other things being equal, fixing a likelihood of *p* sufficiently bigger than 0 and smaller than or equal to 0.5, we expect *wh-*QUD to be associated with comparatively higher choice rates of *probably p* and polar QUDs to be associated with comparatively higher choice rates of *probably not p.* This translates into a hypothesis about our regression model (see below), namely that the coefficient for (non-reference level) QUD:WH should be credibly different from and if fact higher than 0.

**2. Sampling plan**

**- existing data**

*(Preregistration is designed to make clear the distinction between confirmatory tests, specified prior to seeing the data, and exploratory analyses conducted after observing the data. Therefore, creating a research plan in which existing data will be used presents unique challenges.)*

Three preliminary versions of the present study have been run, in order to make sure that the overall design and its implementation worked as expected. None of the pilot versions shared the exact same design and/or measures with the present study.

**- explanation of existing data**

*(If you indicate that you will be using some data that already exist in this study, please describe the steps you have taken to assure that you are unaware of any patterns or summary statistics in the data. This may include an explanation of how access to the data has been limited, who has observed the data, or how you have avoided observing any analysis of the specific data you will use in your study. The purpose of this question is to assure that the line between confirmatory and exploratory analysis is clear.)*

No statistical test has been run on previously collected data, nor have any summary statistics been calculated. No previously collected data will be used in the present study.

**- data collection procedures**

*(Please describe the process by which you will collect your data. If you are using human subjects, this should include the population from which you obtain subjects, recruitment efforts, payment for participation, how subjects will be selected for eligibility from the initial pool (e.g. inclusion and exclusion rules), and your study timeline.)*

Participants with IP addresses in the USA and approval rate of at least 90% will be recruited on Amazon Mechanical Turk (AMT). Participants will be paid 2.00 USD for their participation. The study will run on AMT for 24h or until the desired sample size is obtained.

**- sample size & sample size rationale**

*(Describe the sample size of your study. How many units will be analyzed in the study? This could include a power analysis or an arbitrary constraint such as time, money, or personnel.)*

50 participants will be recruited on AMT. This number should suffice for stable posterior inference of model coefficients (see below), but might not be enough to allow for trustworthy model comparison with LOO, in which case we resort to K-fold cross-validation. Cost reasons prevent us from increasing the number of participants.

**- stopping rule**

*(If your data collection procedures do not give you full control over your exact sample size, specify how you will decide when to terminate your data collection.)*

Not applicable.

**3. Variables**

**- manipulated variables**

*(Describe all variables you plan to manipulate and the levels or treatment arms of each variable.)*

Each research question/hypothesis as specified above corresponds to exactly one manipulated variable.

1. VALUE (metric): the likelihood of a certain event *p* (the focal outcome) to obtain is manipulated. This concept is operationalized as the ratio of balls of a certain color contained in an urn. Specifically, in each trial participants will be shown a picture depicting an urn containing 10 colored balls, *n* of which will be of a specific color, for example red. The ratio *n*/10 corresponds to the likelihood of the event ‘a ball randomly drawn from the urn will be red’ to obtain. The number *n* will have possible values in the set {0,3,4,5,10}. Trials with *n=0* or *n=10* will be considered control trials for the purpose of exclusion of participants. Control trials will also enter the regression analyses.
2. SCENARIO (binary factor, unordered, levels: ‘dual’ (reference level) and ‘plural’): the configuration of the event space is manipulated. In each trial, the content of the urn displayed to participants will be displayed in either of two possible configurations, referred to as *dual* and *plural*. In the *dual* configuration, the balls in the urn will be of exactly two colors: for example, *n* balls will be red and 10-*n* will be blue. In the *plural* configuration, the balls in the urn will be of more than two colors: for example, *n* balls will be red and 10-*n* will be each of a different color.
3. QUD (binary factor, unordered, levels: ‘polar’ (reference level) and ‘wh’): the question under discussion in the conversation is manipulated. In each trial participants will have to send a message in order to answer a fictitious collaborator’s question about the content of the urn. The question will either be a *wh*-question about the color of a randomly drawn ball (e.g. ‘Which color will a randomly drawn ball be?’) or a *polar* question about a specific color (e.g. ‘Will a randomly drawn ball be red?’).

In addition to these, the color of the focal outcome will be manipulated (red, blue, white, black). No research question/hypothesis is associated with this manipulation, whose goal is to make the task more entertaining for the participants and to avoid potential effects of color focality.

**- measured variables**

*(Describe each variable that you will measure. This will include outcome measures, as well as any predictors or covariates that you will measure. You do not need to include any variables that you plan on collecting if they are not going to be included in the confirmatory analyses of this study.)*

In each condition, participants will have to choose a probability expression to best describe the situation to a fictitious collaborator. Possible choices will always include exactly four expressions, namely *certainly not*, *probably not*, *probably*, *certainly*. We record participants’ choices of probability expressions in each condition. As choice options are naturally ordered by logical strength (*certainly not <* *probably not <* *probably <* *certainly*) we will treat the measured variable EXPRESSION as count data on an ordinal measure.

**- indices**

*(If any measurements are going to be combined into an index (or even a mean), what measures will you use and how will they be combined? Include either a formula or a precise description of your method. If your are using a more complicated statistical method to combine measures (e.g. a factor analysis), you can note that here but describe the exact method in the analysis plan section.)*

Not applicable.

**4. Design plan**

**- study type**

*(Please check one of the following statements)*

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

**- blinding**

Not applicable.

**- study design**

*(Describe your study design. Examples include two-group, factorial, randomized block, and repeated measures. Is it a between (unpaired), within-subject (paired), or mixed design? Describe any counterbalancing required. Typical study designs for observation studies include cohort, cross sectional, and case-control studies.)*

Participants will complete two phases. The first phase is a familiarization task, where participants will be introduced to a fictitious two-player betting game in which one player (referred to as the ‘receiver’) can ask a hint about the content of an urn to another player (‘sender’), and then decide if/how he/she wants to place a bet about the content of the urn. In the first phase participants will play three rounds in the role of receiver. In each round, participants will be given a betting option (e.g. ‘You can bet on red or nor bet at all’) and will have to ask a question to the sender, choosing between a *wh*-question or a polar question. After having received an answer from the sender (in the form of a sentence containing a probability expression such as the ones investigated in this study), participants will have to decide if/how they want to bet.

The second phase is the main experimental task, in which the three variables (VALUE, SCENARIO, QUD) will be manipulated, as described above. Participants will complete sixteen trials in the role of sender, i.e. they will read a question asked by another fictitious player, observe the content of an urn and send a message choosing a probability expression. This phase will consist of twelve critical trials, interspersed with four control trials. In each of the critical trials, participants will see one of the possible twelve combinations of conditions obtained crossing three non-trivial quantities of balls of the focal color (VALUE: 3, 4, 5) with two SCENARIOs (plural, dual) and two QUDs (*wh*, polar). (See the attached example stimuli ‘polar-dual.png’, ‘polar-plural.png’, etc.) Each participant will see each combination exactly once. In the control trials, participants will see exclusively trivial quantities of balls of the focal color, i.e. VALUE 0 or 10, together with random selection of SCENARIO and QUD.

**- randomization**

*(If you are doing a randomized study, how will you randomize, and at what level?)*

The twelve critical conditions will be presented to participants in random order.

**5. Analysis plan**

**- statistical models**

*(What statistical model will you use to test each hypothesis? Please include the type of model (e.g. ANOVA, multiple regression, SEM, etc) and the specification of the model (this includes each variable that will be included as predictors, outcomes, or covariates). Please specify any interactions that will be tested and remember that any test not included here must be noted as an exploratory test in your final article.)*

We take a Bayesian data analysis approach to ordinal mixed effects regression models using R and the ‘brms’ package (see attached script for more details). The target model whose coefficients’ posterior distribution we are interested in regresses the ordinal factor EXPRESSION (choice frequencies of probability expression) against VALUE, SCENARIO and QUD and all of their potential interactions, as well as the maximal random effects structure (by-participant random intercepts and slopes for all explanatory variables). We test our hypotheses about the influence of manipulated variables by checking whether the estimated posterior mass of coefficients for main effects of VALUE, SCENARIO and QUD is credibly (in terms of 95% highest density intervals) different from 0 and goes in the expected direction (see above under “Hypotheses”). We rely on the default prior assumptions of the ‘brms’ package for simplicity.

In the unexpected event that the maximal random effects structure turns out to be too complex so that the MCMC sampling does not converge in reasonable time we will reduce it by dropping random effects, prioritizing the removal of interaction terms and prioritizing the removal of explanatory factors in the order first QUD, then SCENARIO, then VALUE, and finally also dropping the by-subject random intercept.

To further test our hypotheses about the influence of factors VALUE, SCENARIO and QUD we will also compare the target model to alternative simpler models. (See script ‘mock\_analysis.R’ for full specification.) In particular, we will look at models that drop either one of the three explanatory factors, that keep just one of them and a null intercept-only model. In each case, we include all possible interactions and the corresponding maximal random effects structure. In case of non-convergence, we will use the same procedure as described above to trim the random effects structure.

We will compare these models using the PSIS-LOO scores computed via the ‘loo’ package. We are specifically interested in the question whether dropping explanatory factors VALUE, QUD or SCENARIO generally leads to an increase in LOO-ICs (equivalently ELPMs). If so, this is evidence that the factor is relevant for the (posterior) predictive success of a model.

If model comparison by LOO is infeasible because of too-high Pareto-smoothing constants *k* (as detailed in Vehtari, Gelman & Gabry, 2015), we resort to 7-fold cross-validation. In this case, we will use a random selection of 2 data points from each participant to form a 7-fold partition of our data and compute expected log point-wise probability mass manually. In this case, we will not be able to use standard error estimates and will rely even less on the results of model comparison, treating them only as a sanity check.

We will compare the results from parameter inference and model comparison. Ideally, results from both approaches would agree. But this might not be the case. If so, we prioritize results from parameter inference and discuss potentially diverging signatures in the light of the conceptually different perspectives taken by inference and model comparison.

**- transformation**

*(If you plan on transforming, centering, recoding the data, or will require a coding scheme for categorical variables, please describe that process.)*

No transformation of the data is planned.

**- follow up analyses**

*(If not specified previously, will you be conducting any confirmatory analyses to follow up on effects in your statistical model, such as subgroup analyses, pairwise or complex contrasts, or follow-up tests from interactions? Remember that any analyses not specified in this research plan must be noted as exploratory.)*

**- inference criteria**

*(What criteria will you use to make inferences? Please describe the information you’ll use (e.g. specify the p-values, Bayes factors, specific model fit indices), as well as cut-off criterion, where appropriate. Will you be using one or two tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?)*

As for parameter inference, we will use 95% highest density intervals to check for inclusion of coefficient value 0. We will not assume “regions of practical equivalence” (Kruschke, 2016) but use visualization and give a critical interpretation of this hard cut-off criterion if feasible. As for model comparison, we will use the estimated standard errors included in the ‘loo’ package to critically assess whether any differences in ELPMs are substantially different.

**- data exclusion**

*(How will you determine which data points or samples (if any) to exclude from your analyses? How will outliers be handled?)*

We will exclude all data points obtained from participants who report a native language different from English, and all data points obtained from participants who failed 3 or more control trials (out of 4). If participants express technical problems or lack of understanding of the task in an optional comments field, we will also discard their data from the analysis. Should technical issues occur so that data points are missing from a single participant, we will also exclude the data from this participant.

**- missing data**

*(How will you deal with incomplete or missing data?)*

Not applicable.

**- exploratory analysis**

*(If you plan to explore your data set to look for unexpected differences or relationships, you may describe those tests here. An exploratory test is any test where a prediction is not made up front, or there are multiple possible tests that you are going to use. A statistically significant finding in an exploratory test is a great way to form a new confirmatory hypothesis, which could be registered at a later time.)*

No further exploratory analysis is planned.

**5. Scripts**

*(This optional step is helpful in order to create a process that is completely transparent and increase the likelihood that your analysis can be replicated. We recommend that you run the code on a simulated dataset in order to check that it will run without errors.)*

The attached script “mock\_analysis.R” contains the main steps of our analysis on a mock data set, available as ‘simdata.csv’, which we created by carelessly clicking through the experiment ourselves 10 times and taking 5 copies of each recorded set of results.