Replication of Experiment 1 of “The Adams family”

(Douven & Verbrugge 2010, Cognition, 117:302-318)

Background

Adams (1975) famously proposed that the assertability of a conditional sentence $P \rightarrow Q$ is given by the conditional probability $P(q \mid p)$ on some probability distribution representing the uncertainty of the speaker or of the interlocutors' combined. Jackson (1975) proposed a reformulation, namely that the **acceptability** of a conditional sentence $P \rightarrow Q$ is given by the conditional probability $P(q \mid p)$. We call this idea Adams' thesis (AT).

Adams’ thesis has been influential in philosophy and theoretical linguistics, and so Douven & Verbrugge (2010) tested the empirical adequacy of AT. Their Experiment 1 directly tests one interpretation of AT and one experimental operationalization of the theoretical notions involved. We seek to replicate their Experiment 1.

Notable differences between original and replication

Major differences between original and replication are:

- the original was a paper-based with university students who did the study for course credit; the replication uses paid online recruitment (via Prolific)
- the original was conducted in Dutch; the replication uses the English translations of the original material (as offered in the appendix of the original paper)
- we replace frequentist ANOVAs from the original paper with Bayesian hierarchical regression

Hypotheses to be tested

While Douven & Verbrugge (2010) considered even more variation in the hypotheses to be investigated, we focus here on the following two main research hypotheses, each of which has a sub-ordinate hypothesis concerning the differential effects of different types of conditions (as discussed in the original paper):

- **H1-AT:** Since according to a strong reading of AT, there is no difference between “acceptability” and “conditional probability” of a conditional, we expect is no difference between ratings given for the "acceptability" and those given for the "conditional probability" condition in the experiment described below, i.e., we expect a main effect of the kind of rating subjects gave.
  - **H1-types:** We expect all types of conditionals tested (deductive, inductive and abductive; as introduced in the original paper; see below), to behave the same; i.e., we expect no main effects of or interactions based on types of conditional sentences.
- **H2:** Since according to a weaker reading of AT, there should at least be a correlation between “acceptability” and “conditional probability”, we expect that, when looking at the averages for each vignette (see below), ratings of "acceptability" should correlate positively with those of "conditional probability”.
  - **H2-types:** We expect no effects of the type of conditional on the correlation.

Methods

The experiment is presented as a dynamic browser app, using magpie. The experiment can be inspected here.

Design

We use a factorial design with two factors as **independent variables**:

1. RATING: is a between-subjects factor with unordered levels ("acceptability" and "conditional probability") corresponding to the kind of question the participants answered; we treat acceptability as the default level in dummy-coding (see below)
2. COND-TYPE: is a within-subjects factor with unordered levels ("deductive", "inductive", "abductive") corresponding to the kind of conditional sentence that was presented; we will use "deductive" as the default level in dummy-coding (see below)

The **dependent measure** is a choice on a 7-point Likert scale. Despite the known issues with this interpretation, we follow the original paper in treating this data as metric; e.g., we compute averages over ratings to address the predicted correlation in H2 and H2-types, as in the original paper.

Planned sample

Participants for this study will be recruited via the crowd-sourcing platform Prolific. Eligible are participants who self-identify as native speakers of English (any variety) and who have at least a 90% acceptance rates on previous experiments on Prolific.

The intended sample size is 160 participants, thus more than doubling the sample size of the original paper (N=67) and resulting in approximately 80 participants in each group (by random group allocation). Online recruitment via Prolific can occasionally produce data from more participants than requested, in which case we will use the first 160 participants only, but would use any surplus to substitute any excluded participants. We will not re-recruit in case of participant exclusion otherwise.

Exclusion criteria

We will exclude all data from each participant who responded in less than 3 seconds on more than 2 trials. We will exclude all individual trials with response times below 3 seconds. This is because we judge it impossible to properly read the material in less than 3 seconds.

Materials

We will use the English translations of all original materials, as provided in Appendix A of Douven & Verbrugge's (2010) paper.

There are 30 vignettes. Each vignette has a context and a (conditional) statement. There are 10 vignettes for each kind of conditional.

Procedure

Participants first receive general instructions about their rights and the use of the data. They then see specific instructions explaining the task to them.

Each participant is initially randomly assigned to a group, determining the context and statement variants that they see. Each participant then goes through all 30 vignettes, in a completely randomized order (determined on-the-fly at the beginning of the experiment).

After the 30 trials have been completed, participants can leave additional information on age, gender, education and any comments they might have regarding the experiment.

Data preprocessing

Data will be cleaned only by following the above mentioned exclusion criteria.

For each vignette, we will look at the average rating of “acceptability” and of “conditional probability”, that is averaged over all participants. This is in order to address H2 and H2-types, in line with the original paper.

Statistical models

An example data analysis file, based on pilot data from 10 participants (which will not be included in the final data analysis), can be found HERE (ADD LINK).

Hypotheses H1 and H1-types

To test hypotheses H1 and H1-types, we will use Bayesian regression analysis, as implemented in the R package “brms” (Bürkner 2016). In keeping with the original paper’s treatment of the data as metric, we first regress rating scale choices as numeric responses (1-7) against regressors RATING, COND-TYPE and their interaction. We also add the random intercepts for both subjects and items, but refrain from making the model more complex (for the sake of simplicity, in the context of this example preregistration).

RESPONSE ~ RATING * COND_TYPE + (1 | subject_id + vignette)

We will use weakly conservative Gaussian priors (e.g. Gelman 2006) centered around 0 (sd = 10) for all population-level regression coefficients.

Hypotheses H2 and H2-types

The averaged ratings for each vignette (N=30) will be subjected to a linear regression. We will regress averages of “acceptability” on “conditional probability”, because it seems most natural to think of “conditional probability” as the source of “acceptability” ratings, from the theoretical point of view we focus on here. To test H2, we look at all the (aggregate) data and look at the model:

ACCEPTABILITY ~ COND_PROB

To test H2-types, following the original paper, we look at the same model, ran separately three times, once for each type of conditional sentence.

We use a weakly conservative Gaussian prior (mean 0, sd = 10) for regression coefficients for COND_TYPE and its interaction, and a weakly conservative Gaussian prior (mean = 1, sd = 10) for the regression coefficient for COND_PROB.

Inference criteria

We report the posterior probability of regression coefficients.

To test H1, we look at the posteriori probability that mean judgements of “conditional probability” (across all types of conditionals) is credibly different from the mean judgements of “acceptability”. We consider H1 to be discredited by the data if the posterior probability that the

difference in means is bigger than zero exceeds .95 or is lower than .05. We judge H1-types to be discredited if, by the same measure as for H1 but for each conditional type separately, not all conditionals behave the same.

To test H2 and H2-types, we look at the posteriori probability that the regression coefficient for COND_PROB (the slope of the linear regression) is smaller than zero. We consider H2 to be discredited if the posterior probability that the slope is positive is smaller than 0.95. We consider H2-types to be discredited if, by the same criterion as for H2, we would reach different conclusions for different types of conditionals.