Supplemental Material

PT & CS

Contents

1 Contingency Awareness **3**

1.1 Experiment 1 . . . 3

1.2 Experiment 3 . . . 3

1.3 Reliability . . . 4

2 Inclusion Bayes Factors for Key-Switch, Target-Type, and Valence **5**

2.1 Motor Response Learning . . . 5

2.2 Perceptual Learning . . . 5

2.3 Evaluative Learning: Ratings . . . 6

2.4 Evaluative Learning: Affect Misattribution Procedure . . . 6

3 Awareness subgroup comparison: Selection based on even-numbered trials **7**

3.1 Perceptual Awareness . . . 7

3.2 CS Discrimination . . . 8

4 Estimating best-fitting parameter values for the predictor data **10**

4.1 Selecting simulation parameters . . . 10

4.2 Determining the magnitude of measurement error of X . . . 10

4.3 Reliability . . . 11

4.4 Distribution of Error . . . 11

4.5 Goodness of fit . . . 11

4.6 Obtaining the true Y values . . . 12

4.7 Results . . . 12

5 EiV Robustness Checks, Part 1 **13**

5.1 Rationale . . . 13

5.2 Model . . . 13

5.3 Results . . . 15

5.4 Slope Power - Standard Regression . . . 20

6 EiV Robustness Checks, Part 2 (SDT Model) 25

6.1 Rationale . . . 25

6.2 Model . . . 25

6.3 Main Results . . . 28

6.4 Additional Results . . . 35

7 EiV Robustness Checks, Part 2 (Normal Measurement Error) 39

7.1 Rationale . . . 39

7.2 Model . . . 39

7.3 Main Results . . . 40

7.4 Additional Results . . . 40

1 Contingency Awareness

We additionally assessed contingency awareness in Experiments 1 and 3; the results of this measure were however not reported in the paper because the measure was unreliable. The measure is described here for completeness.

1.1 Experiment 1

Participants were instructed that, in some experimental conditions, CS and US presentations followed a contingency, and they were asked to indicate this contingency (or, if they did not notice a contingency, to use their intuition or guess). For every CS and distractor, participants indicated whether it appeared "before pleasant words (certainly)", "before pleasant words (probably)", "before pleasant words (guess)", "before unpleasant words (certainly)", "before unpleasant words (probably)" or "before unpleasant words (guess)". All items were presented on one screen, the first two were always distractors, and the order of the remaining items was pseudo-random.

On average, participants correctly identified the valence of the two CSs not more often than chance (chance level = 1 CS), Exp.1A: $M = 1.00$, 95% CI $[0.85, 1.15]$, $t(84) = 0.00$, $p > .999$, Exp. 1B: $M = 0.95$, 95% CI $[0.79, 1.11]$, $t(84) = -0.59$, $p = .558$. When responses were coded on a scale ranging from –2.5 (indicating incorrect valence and high certainty for both CSs) to +2.5 (indicating correct valence and high certainty for both CSs), the mean awareness score was also not better than chance (chance level = 0), Exp. 1A: $M = 0.09$, 95% CI $[-0.13, 0.31]$, $t(84) = 0.79$, $p = .429$, Exp. 1B: `ttest_awa2_exp1B`.

Note that our scale differed from that used by G&DH. In the original paradigm, the middle response option was "[CS] did not appear in the task". We found this problematic as it is not a natural zero point of the scale; it is not an answer to the question of whether a CS appeared before pleasant or unpleasant USs, but to the question of whether the CS appeared at all. Hence, we removed the "did-not-appear" item from the scale and forced participants to decide whether they rather thought that a CS appeared before pleasant or unpleasant words. Because of the scale's ordinal nature, we refrained from using the regression method and instead compared subgroups of participants who reported the correct valence for none, one, or both CSs. By assessing contingency awareness before the conditioning test, we ensured that the contingency awareness test reflects the knowledge about the critical contingency present in the acquisition phase (instead of the 50% contingency in the conditioning test, which is equivalent to an extinction procedure).

1.2 Experiment 3

Contingency awareness was tested as in Exp. 1, but as for the evaluative rating, there were only two distractors. The first item was a distractor and the order of the remaining items was pseudo-random. The contingency awareness test applied by G&DH and in Exp. 1 explicitly asks whether the CSs were paired with pleasant or unpleasant USs. However, the learning effect may not be evaluative but instead rely on non-evaluative knowledge (e.g., semantic knowledge, perceptual information such as letters or their features) that is consciously accessible but not captured by the contingency awareness test. In Exp. 3, the contingency-awareness instructions listed the USs and informed participants about the valence categories to which the USs belonged, so that they could use these categories to adequately report their knowledge in the subsequent contingency assessment.

On average, participants correctly identified the valence of the two CSs not better than expected by chance (chance level = 1 CS), $M = 0.98$, 95% CI $[0.87, 1.09]$, $t(169) = -0.43$, $p = .671$. When responses were coded on a scale ranging from –2.5 (indicating incorrect valence and high certainty for both CSs) to +2.5 (indicating correct valence and high certainty for both CSs), the mean awareness score was also not better than chance (chance level = 0), $M = 0.04$, 95% CI $[-0.12, 0.19]$, $t(169) = 0.46$, $p = .645$.

1.3 Reliability

Contingency awareness was assessed in Experiments 1 and 3. The critical contingency is between the identity of the CS and the valence of the USs paired with that CS (i.e., the response category). Participants reported, for each of the two CSs, the valence of the USs paired with that CS. We treated these two contingency ratings (i.e., one for each CS) as independent measures of contingency awareness. To estimate the reliability of the contingency awareness measure, we computed interrater agreement (Cohen's Kappa). Results indicated practically zero reliability in Exp. 1 (70 ms: κ = -.04; 80ms: κ = .02), suggesting that the measure does not provide any information (i.e., reflects random noise), and very low reliability in Exp. 3 (κ = .19), suggesting too little systematic variance worthy of further analysis.

2 Inclusion Bayes Factors for Key-Switch, Target-Type, and Valence

In addition to the BFs based on meta-analyses reported in the paper, we also ran Bayesian ANOVAs and computed *Inclusion BFs* to determine the amount of evidence against an effect of our experimental manipulations (Key-Switch, Target Type, US Valence).

2.1 Motor Response Learning

In Experiments 1-3, we tested whether the conditioning effect was affected by response key assignment (i.e., whether it was smaller when response keys were switched after the learning phase). To probe whether the null effects reported above represent evidence for the absence of motor contributions to the effect, we compared a null model with the factors Experiment and Order (evaluative measures vs. conditioning test first) to an alternative model that also included the Key Assignment factor. Both models also included the interaction between Experiment and Order, but no interactions involving Key Assignment (i.e., Type 2 sums of squares). To avoid nested factors in the joint analyses, we included only those conditions that are repeated across all studies (i.e., we excluded the new-targets and 70ms conditions). [1] We report robustness regions of the Bayes factor giving the minimum and maximum scale value of the prior where $BF_{10} < 1/3$ or $BF_{10} > 3$ [@dienes_how_2019]. The prior indicates the scale parameter of a Cauchy distribution with location = 0 representing the assumptions about the expected effect size, a small prior assumes that the effect size is small and a large prior assumes a larger effect size. A $BF_{10} < 1/3$ indicates moderate (or stronger) evidence for the null against the alternative model, and a $BF_{10} > 3$ for the alternative against the null model.

The model comparisons yielded evidence for the null model with medium and wider priors, $RR_{BF<1/3}[.35, \infty]$ (for smaller priors the BF was inconclusive), e.g. there was evidence against medium and larger effects of Key Assignment on the conditioning effect. Consistent with our interpretation of the individual experiments, the conditioning effect is unlikely to be exclusively due to motor learning. Yet, based on the present data we cannot rule out a small motor contribution.

Parallel to the frequentist ANOVA models reported for the individual experiments, we also performed a Bayesian analysis with both the null and the alternative model including all two- and three-way interactions with Key Assignment. Results confirmed the evidence against medium and large effects of Key Assignment on the conditioning effect, $RR_{BF<1/3}[.29, \infty]$ (for smaller priors the BF was inconclusive).

2.2 Perceptual Learning

In Experiments 2 and 3, we investigated whether conditioning involved perceptual learning about the USs by comparing conditioning effects for US targets with those for new target stimuli. To analyse whether perceptual learning contributed to the conditioning effect, we compared a null model with the factors Experiment, Order (evaluative measures vs. conditioning test first), and Key Assignment with an alternative model that additionally included the factor Target Type. Both models included by-subject random effects, as well as the three-way and all two-way interactions between Experiment, Order, and Key Assignment, but no interactions involving Target Type. We again tested a directed hypothesis (expecting a larger conditioning effect with USs vs. new targets).

There was evidence against an effect of Target Type on the conditioning effect even for very small priors, $RR_{BF<1/3}[.11, \infty]$ (for even smaller priors the BF was inconclusive). Thus, results strongly suggest that perceptual learning does not play any relevant role in the conditioning effect.

[1] We used the lmBF function from the BayesFactor package [@R-BayesFactor]. In this and the following analyses reported in this section, the prior on all irrelevant effects was set to $r = .1$ because only small if any effects were expected. We tested a directed hypothesis (expecting a larger conditioning effect with same vs. switched keys) by sampling from the posterior of the full model and multiplying the BF for the full model against the null model with the relative frequency of hypothesis-consistent samples in the posterior, then divided it by 0.5 (i.e., the prior odds of obtaining a larger conditioning effect for same vs. switched keys).

As for the motor account, we also compared ANOVA models that were more comparable to those of the frequentist analyses reported above, with both the null and the alternative model including all two- three-, and four-way interactions with Target Type. There was evidence against an effect of Target Type on the conditioning effect only for medium and wider priors, $RR_{BF<1/3}[.47, \infty]$ (for smaller priors the BF was inconclusive).

2.3 Evaluative Learning: Ratings

To probe for evaluative learning effects, we administered both evaluative ratings and the AMP task. Here we analyzed evaluative ratings of the CSs from Experiments 1-3 to more conclusively determine whether or not ratings were more positive for CS_{pos} than for CS_{neg}. [2] To analyze effects of US valence on evaluative ratings, we compared the null model with the factors Experiment and Order (evaluative measures vs. conditioning test first) with the alternative model that additionally included the Valence factor. Both models also included by-subject random effects, as well as the two-way interaction between Experiment and Order, but no interactions involving Valence. We tested a directed hypothesis (expecting more positive evaluative ratings with CS_{pos} than with CS_{neg}).

The simpler null model was preferred in model comparisons, yielding evidence against an evaluative rating effect even for very small priors, $RR_{BF<1/3}[0.09, \infty]$ (for even smaller priors the BF was inconclusive). These findings provide strong support for the absence of evaluative learning in the G&DH conditioning effect.

To parallel the frequentist ANOVAs of the individual experiments, we also performed a Bayesian analysis with both the null and the alternative model including all two- and three-way interactions with Valence. Results confirmed evidence against a Valence effect even for small priors, $RR_{BF<1/3}[.24, \infty]$ (for smaller priors the BF was inconclusive).

2.4 Evaluative Learning: Affect Misattribution Procedure

Finally, we jointly analyzed the indirect evaluations obtained from the AMP task in Experiments 2 and 3, in which participants classified unfamiliar ideographs as pleasant or unpleasant. The ideographs always followed the presentation of a CS, and an evaluative learning effect would be reflected in more pleasant ratings after the CS_{pos} than after the CS_{neg}.

We compared the null model, containing the factors Experiment, Order (evaluative measures or conditioning test first), and Key Assignment, to an alternative model that had an additional Valence factor. Both models also included by-subject random effects, as well as all two-way and the three-way interaction between Experiment, Order, and Key Assignment, but no interactions involving Valence. We tested a directed hypothesis (expecting more "positive" responses after the CS_{pos} than after the CS_{neg}).

There was evidence against an AMP effect only for very large priors, and for all smaller priors the BF was inconclusive, $RR_{BF<1/3}[1.02, \infty]$. This shows that, despite the non-significant effects, we cannot rule out small, medium or even large effects in the AMP. The alternative Bayesian analysis even yields evidence for an AMP effect.

We also compared ANOVA models more comparable to our above frequentist analyses, with both the null and the alternative model including all two-, three-, and the four-way interaction with Valence. For this type of analysis, the BF was inconclusive over the complete range of tested priors between $r = 0.1$ and $r = 1.5$.

[2] As noted above, to avoid nested factors we excluded the 70ms condition of Exp.1.

3 Awareness subgroup comparison: Selection based on even-numbered trials

In the manuscript, we used the odd-numbered trials to define low- and high-awareness groups, and described their awareness levels based on the even-numbered trials as an independent measure. Using this odd-numbered classification, we then compared conditioning between the low- and high-awareness groups. Since the choice of odd- or even-numbered trials for group classification is arbitrary, we now present the equivalent analyses using the even-numbered trials to define the awareness groups. For clarity and ease of comparison, we have retained the same textual structure as in the original description, modifying it only where necessary.

```
## `summarise()` has grouped output by 'exp'. You can override using the `.groups`
## argument.
```

3.1 Perceptual Awareness

After creating these subgroups, we used the second measure (based on the **odd**-numbered trials) as an unbiased assessment of the subgroups' level of perceptual awareness. Descriptively, perceptual awareness was close to zero in the low-awareness subgroups, in which the proportion of PAS3 or PAS4 ratings varied between 0 and .17, and $\leq$ 23% of participants ever reported perceiving even a single letter. Awareness was substantially higher in the high-awareness subgroups, with PAS3 or PAS4 ratings in approximately 1 out of 4 trials (.17 - .37) and 92-100% of participants perceiving one or more letters at least once.

For comparisons across subgroups and between measures, we transformed participants' proportions of (odd-numbered) trials with PAS3 or PAS4 ratings into Cohen's standardized effect size measure h [@cohen_statistical_1988; conventional small, medium, and large effects are $h = .2$, $h = .5$, and $h = .8$]. Figure @ref(fig:evenodd-plots) (upper left panel) presents mean perceptual-awareness effects by subgroup for each experiment, along with a meta-analytic mean (based on Bayesian model-averaging as outlined above). Perceptual awareness levels robustly differed between low- and high-awareness subgroups (i.e., the CI around their difference did not include zero, $\Delta h = 0.94$, 95%CI [0.65, 1.22], and the difference was significant in all four studies, all ps $< .001$).

Perceptual awareness in the *low* subgroups did not differ from zero (i.e., the CI around the meta-analytic mean included zero, $h = 0.08$, 95% CI $[-0.02, 0.18]$, corresponding to 0.1% PAS34 ratings), and awareness did not differ from zero in any of the four studies, all ps $\geq .149$). We computed Bayes Factors to assess the evidence for the null hypothesis of zero Perceptual Awareness in the low-awareness subgroups. BFs remained inconclusive for (very) small effect sizes, but indicated moderate evidence against small-to-medium effect sizes of Cohen's $h > 0.35$ and strong evidence against **large** effect sizes of Cohen's $h > 0.86$. Perceptual awareness in the *high* subgroups robustly differed from zero (i.e., the CI around the meta-analytic mean did not include zero, $h = 1.01$, 95% CI $[0.74, 1.28]$, corresponding to 23% PAS34 ratings), and awareness was different from zero in all of the four studies, all ps $< .001$).

Given this difference in awareness between subgroups, we probed whether the subgroups also differed in their conditioning effect. For comparability with the awareness effect, we computed Cohen's h as a measure for how strongly participants' proportion of congruent responses deviated from chance. The upper middle panel of Figure @ref(fig:evenodd-plots) shows mean conditioning effects as a function of awareness subgroup for each experiment. Conditioning effects did not differ across awareness levels (i.e., the CI around their difference included zero, $\Delta h = 0.00$, 95%CI [-0.02, 0.02], and the difference was not significant in any of the experiments, all ps $\geq .210$). Bayes Factors indicated moderate (strong; very strong) evidence against small effect sizes of $h > 0.02$ (0.04; 0.09). Overall, results yielded conclusive evidence that conditioning did not vary as a function of Perceptual Awareness subgroup.

Conditioning effects were robust at both levels of awareness. In the low-awareness subgroups, the conditioning effect was significant (i.e., the CI around the mean conditioning effect did not include zero, $h = 0.03$, 95% CI $[0.02, 0.05]$; conditioning was significant in experiments 1B and 3, $p \leq .001$, **but not in Exp. 1A**,

p =.063, and Exp. 2, p =.133). The high-awareness subgroups also showed a significant conditioning effect over all experiments (i.e., the CI around the mean conditioning effect did not include zero, $h = 0.03$, 95% CI $[0.03, 0.04]$ (both subgroup estimates correspond to 52% CS consistent responses), and conditioning was significant in all four experiments, all $p \leq .007$).

The upper right panel of Figure @ref(fig:evenodd-plots) shows Conditioning effect sizes against Perceptual Awareness effect sizes (both in Cohen's h units). The plot shows the subgroup means (leftmost points: low-awareness; rightmost points: high-awareness), with lines connecting the points belonging to the same Experiment and a bold line representing the meta-analytic mean. **In case of selection based on even-numbered trials, all points lie on or below the main diagonal. Thus, conditioning (the indirect effect) was never larger than awareness (the direct effect), and there is no evidence for an indirect task advantage.**

3.2 CS Discrimination

We again used a binomial test to detect above-chance CS discrimination, now based on even-numbered instead of odd-numbered trials. Participants with above-chance CS discrimination were assigned to the high-awareness subgroup. The remaining participants were assigned to the low-awareness subgroup.

Awareness levels of these subgroups were then assessed using the unbiased second measure (i.e., based on odd-numbered trials). Figure @ref(fig:evenodd-plots) (lower left panel) illustrates that CS discrimination robustly differed across subgroups (i.e., the CI around their difference did not include zero, $\Delta h = 0.22$, 95%CI [0.06, 0.39], and the difference was significant in 3 out of 4 experiments, all ps $< .001$ except Exp.1A, $p = .129$). In the low-awareness subgroups, CS discrimination was descriptively just above zero, but the CI around the mean included zero, $h = 0.07$, 95% CI $[-0.005, 0.14]$, corresponding to 53% correct CS classifications; discrimination was not significant in 3 out of 4 experiments, all ps $\geq .108$ **except Exp. 2**, $p = .046$). **Bayes Factors yielded evidence for the alternative hypothesis of above-chance-level CS Discrimination in the low-awareness subgroups for small-sized effects** (moderate evidence for $h \geq 0.01$ and ≤ 0.37, strong evidence for $h \geq 0.02$ and ≤ 0.17, and very strong evidence for $h \geq 0.04$ and ≤ 0.08). For medium-to-large effects the Bayes Factors were inconclusive. In the high-awareness subgroups, CS discrimination increased to $h = 0.29$, 95% CI $[0.14, 0.44]$, corresponding to 64% correct CS classifications, and discrimination was significant in all experiments, all ps $< .001$.

Given this difference in CS discrimination, we probed whether the subgroups also differed in their conditioning effect. Figure @ref(fig:evenodd-plots) (lower middle panel) presents conditioning effects separated by CS Discrimination subgroup. Conditioning effects did not differ significantly across subgroups (i.e., the CI around their difference included zero, $\Delta h = 0.01$, 95%CI [-0.01, 0.03], and the difference was not significant **in any experiment**, all ps $\geq .091$). Bayes Factors indicated moderate evidence for very small effects, $h \leq 0.03$; and moderate (strong) evidence against small effect sizes of Cohen's $h > 0.14$ (0.39). Overall, the evidence suggests that the robust increase in CS discrimination across groups was accompanied by, if anything, a very small increase of conditioning in the high-awareness groups.

The conditioning effect was very small in the low-awareness subgroups (i.e., the CI around the mean excluded zero, $h = 0.03$, 95% CI $[0.01, 0.05]$, corresponding to 51.3% correct CS classifications, and conditioning was **significant in Exp. 1B**, $p = .006$, **and in Exp. 3**, $p = .020$, but not in Exp. 1A, $p = .091$, and in Exp. 2, $p = .299$. Bayes Factors yielded **very strong (strong, moderate) evidence for very small (small, medium) effects**, $h \leq 0.1$ (0.24, 0.7), and was inconclusive for larger effects. Conditioning was **not** larger in the high-awareness subgroups (i.e., the CI around the mean excluded zero, $h = 0.03$, 95% CI $[0.03, 0.04]$, corresponding to 51.7% correct CS classifications, and conditioning was significant in all studies, ps $\leq .005$).

The lower right panel of Figure @ref(fig:evenodd-plots) illustrates how conditioning varies as a function of CS discrimination. All except one subgroup means lie below the main diagonal, indicating that the direct effect (i.e., CS discrimination) was always larger than the indirect effect (i.e., conditioning). A direct comparison of effect sizes for the exception yielded a null result (i.e., the CI around the difference of effect sizes included zero, $\Delta h = 0.01$, 95%CI [-0.13, 0.16]), and Bayes Factors indicated moderate evidence against

an indirect task advantage of more than **small** size (Cohen's $h > 0.41$). As for perceptual awareness, the CS discrimination results thus indicate the absence of an indirect task advantage [@meyen_advancing_2022].

4 Estimating best-fitting parameter values for the predictor data

4.1 Selecting simulation parameters

The goal was to approximate the empirical data obtained our study, using two different types of distributions. The first takes true X values from a truncated-normal distribution, and adds normal measurement noise (with the magnitude determined by a binomial-SDT model). The second takes true X values from a gamma distribution, and adds normal measurement error (with the magnitude again determined by a binomial-SDT model).

For each Experiment, we first searched for the parameters of the true X values that best fitted the observed overall mean and variance of the data.

Next, we obtained the distribution of the true Y values by scaling the X distributions by a factor my/mx (i.e., the slope), and added normal measurement error to match the observed variance of Y.

4.2 Determining the magnitude of measurement error of X

To obtain the best-fitting parameters for X, we used a grid-search approach (search algorithms were not feasible because the measurement error resulting from the binomial-SDT model would have to be estimated at each step of the iteration process as it depends on the true values, and these error estimates are unstable because they have to be computed from random samples). We simplified the task as follows: First, note that the measurement error from the SDT model is determined by the distribution of true sensitivity (d'), so it has no free parameters. Next, we further reduced the problem to a one-dimensional search, based on the notion that the mean of the true-value distribution has to be equal to the mean of the distribution of true values plus error (i.e., it is not changed when adding a random variable with mean of zero). We varied one parameter value and computed the other one based on this equality, so that the mean of the true value distribution was equal to the true-plus-error distribution's mean).

For each point in the grid (i.e., combination of parameter values), we computed the magnitude (variance) of measurement error resulting from the SDT model. We searched for the parameter values for which the sum of true and error variances was closest to the observed variance. Finally, we fine-tuned the parameter values twice by zooming in to the area around the initially selected parameter values and repeating the above approach.

```
##     exp run      mu    sigma mu_rect      diff_mu   sd_rect       bsd    sum_sd
## 26    1   0 -0.1475 1.132807   0.382 2.956141e-09 0.6106434 0.2708730 0.6680251
## 231   2   0 -0.0890 1.718496   0.642 3.107602e-10 0.9727914 0.2929927 1.0159566
## 272   3   0  0.5330 1.009218   0.724 6.174977e-10 0.7589469 0.2828784 0.8099510
## 153   4   0  0.7870 1.440875   1.052 6.319210e-09 1.0909382 0.3138169 1.1351771
##           diff_sd mx_exp sdx_exp
## 26   3.354353e-05  0.382   0.668
## 231 -8.813753e-05  0.642   1.016
## 272 -7.941298e-05  0.724   0.810
## 153  4.021376e-04  1.052   1.135

##     exp run    sc        sh    gm       gsd       bsd    sum_sd       diff_sd
## 38    1   0 0.973 0.3926002 0.382 0.6096606 0.2733271 0.6681270  0.0001697044
## 81    2   0 1.447 0.4436766 0.642 0.9638330 0.3211461 1.0159276 -0.0001472063
## 122   3   0 0.793 0.9129887 0.724 0.7577150 0.2857274 0.8097976 -0.0003278803
## 156   4   0 1.116 0.9426523 1.052 1.0835276 0.3376722 1.1349249 -0.0001704843
##     mx_exp sdx_exp
## 38   0.382   0.668
## 81   0.642   1.016
```

```
## 122  0.724   0.810
## 156  1.052   1.135
```

The resulting parameters are given in the tables above.

4.3 Reliability

We checked whether the amount of measurement error in the best-fitting models can explain the reliability of CS discrimination in the observed data. We sampled observed values from the best-fitting model, added normal measurement error (as an approximation), and computed the squared correlation between observed and true values.

```
##                    Exp.1A     Exp.1B      Exp.2      Exp.3
## R_XX          0.82000000 0.91000000 0.87000000 0.92000000
## R_XX_tn       0.82883872 0.91271734 0.87720002 0.92290278
## sd_R_XX_tn    0.05076489 0.02704804 0.01970511 0.01260047
## R_XX_gamma    0.80947973 0.88428975 0.87061805 0.90816314
## sd_R_XX_gamma 0.07908992 0.05136428 0.03018733 0.02171886
```

The resulting reliability estimates closely match the observed values. This suggests that the model can account well for the predictor data.

4.4 Distribution of Error

We checked whether the distribution of the SDT-generated error after re-transformation into (observed) d' values is approximately normal.

```
##                            Exp.1A    Exp.1B     Exp.2     Exp.3
## sdxerror_tn            0.2708730 0.2929927 0.2828784 0.3138169
## sdxerror_tn_normfit    0.2691653 0.2918160 0.2809591 0.3108307
## sdxerror_gamma         0.2733271 0.3211461 0.2857274 0.3376722
## sdxerror_gamma_normfit 0.2697011 0.2840645 0.2806882 0.3023245
```

The error terms were largely well-fitted by a normal distribution (the KS test rejected the normal fit in ≤ 1.2 of cases). The estimated SD of the normal distribution fitted to the errors closely approximated the error SD obtained above from the model (with the exception of a slight underestimation for the gamma model of Expts. 1B & 3).

4.5 Goodness of fit

Given that the measurement error obtained from the truncnorm-SDT and gamma-SDT models is approximately normally distributed (for the n=96 trials used in our studies), we built on this finding to examine the goodness of fit (AIC) of a truncnorm and a gamma model with normal error (of the magnitude determined above) to the data.

This allowed us to compare the truncnorm and gamma versions of the model.

```
##              Exp.1A    Exp.1B    Exp.2    Exp.3
## nll.tn    46.94098  76.34248 155.3419 213.9837
## nll.gamma 42.90957  74.73813 150.9656 211.7654
## AIC_tn    97.88197 156.68495 314.6837 431.9674
## AIC_gamma 89.81915 153.47627 305.9312 427.5308
```

Results suggest that the gamma distribution consistently yielded somewhat better fits than the truncated-normal.

4.6 Obtaining the true Y values

Using the slope (given by the mean of the empirically observed Y values divided by the mean of the empirically observed X values) as scaling factor, the parameters of the distribution of true Y values can be obtained from the distribution of true X values (they are given in the table below for reference). In the simulations, for each simulated sample we sampled from the distribution of true X values and obtained the true Y values by multiplication with the slope, $Y = \beta * X$

4.7 Results

	Exp.1A	Exp.1B	Exp.2	Exp.3
exp	1	2	3	4
labels	Exp.1A	Exp.1B	Exp.2	Exp.3
nsample	70	68	150	156
ntrialsx	96	96	96	96
my_exp	0.077	0.137	0.065	0.079
sdy_exp	0.216	0.213	0.191	0.164
mx_exp	0.382	0.642	0.724	1.052
sdx_exp	0.668	1.016	0.810	1.135
slope	0.202	0.213	0.090	0.075
mxnorm_tn	-0.1475	-0.0890	0.5330	0.7870
sdxnorm_tn	1.132807	1.718496	1.009218	1.440875
shapex_gamma	0.3926002	0.4436766	0.9129887	0.9426523
scalex_gamma	0.973	1.447	0.793	1.116
sdxerror_tn	0.2708730	0.2929927	0.2828784	0.3138169
sdxerror_gamma	0.2733271	0.3211461	0.2857274	0.3376722
R_XX	0.82	0.91	0.87	0.92
R_XX_tn	0.8288387	0.9127173	0.8772000	0.9229028
sd_R_XX_tn	0.05076489	0.02704804	0.01970511	0.01260047
R_XX_gamma	0.8094797	0.8842898	0.8706181	0.9081631
sd_R_XX_gamma	0.07908992	0.05136428	0.03018733	0.02171886
sdxerror_tn_normfit	0.2691653	0.2918160	0.2809591	0.3108307
sdxerror_gamma_normfit	0.2697011	0.2840645	0.2806882	0.3023245
nll.tn	46.94098	76.34248	155.34186	213.98368
nll.gamma	42.90957	74.73813	150.96558	211.76539
AIC_tn	97.88197	156.68495	314.68372	431.96736
AIC_gamma	89.81915	153.47627	305.93116	427.53077
mynorm_tn	-0.029795	-0.018957	0.047970	0.059025
sdynorm_tn	0.2288269	0.3660397	0.0908296	0.1080656
shapey_gamma	0.3926002	0.4436766	0.9129887	0.9426523
scaley_gamma	0.196546	0.308211	0.071370	0.083700
sdytrue_tn	0.12332450	0.20703863	0.06829620	0.08183213
sdytrue_gamma	0.12325703	0.20531303	0.06821779	0.08126707
sdyerror_tn	0.17733321	0.05004005	0.17837217	0.14212495
sdyerror_gamma	0.1773801	0.0567059	0.1784022	0.1424488

5 EiV Robustness Checks, Part 1

5.1 Rationale

In our simulations, we probed whether the positive intercepts obtained in our data may have been an artifact due to measurement error. The general approach of the simulations was to simulate data from a model that closely resembles the present data but assumes a true zero intercept, to assess by how much standard regression parameter estimates are biased, and to which degree the EiV correction removes the bias and eliminates α-inflation. We used the simulated intercepts as a more appropriate sampling distribution, and computed simulated p-values for our experimental intercepts (how many of the simulated intercepts are at least as positive as the experimental intercept?).

5.2 Model

We assumed a single-process null model with a zero intercept and a perfect correlation between predictor (CS discrimination) and criterion (learning effect). Simulated data were generated by (1) sampling true predictor values (i.e., CS discrimination) from one of two distributions (truncated-normal or gamma), (2) obtaining the true criterion values (i.e., the conditioning effect) from these predictor values based on a zero-intercept model with the slope determined by the data, and (3) to model the measurement error in the predictor by simulating the trials as coming from a binomial distribution and to add normally distributed measurement error to the true criterion values to obtain the observed values.

(1) The parameters of the truncated-normal and gamma distributions used to model the true predictor values ($\mu_{X_{nd}}$ and $\sigma_{X_{nd}}$ of the normal distribution before truncation; shape and scale of the gamma distribution) were obtained by fitting each distribution (plus binomial measurement error determined via a signal-detection model as suggested by Miller, 2000) to the predictor values observed in our studies. Both distributions fitted the observed data very well (see section *Fit and Descriptives*). The resulting best-fitting parameters are given in section *Reliability of the Predictor and True Distributions*.

(2) For the null model of a zero intercept (i.e., no learning in the absence of CS discrimination), we set the intercept to $\beta_0 = 0$. We fixed the slope parameter to $\beta_1 = \bar{y}/\bar{x}$ because, as measurement error does not bias the estimation of the means, the regression line is supposed to always go through $(\bar{x}, \bar{y})$. The true criterion values were generated by multiplying the true predictor values by β_1. (The predicted criterion values were equal to the true criterion values, because the true predictor and criterion correlated perfectly.)

(3) For the predictor, we derived the true probabilities of hits and false alarms (assuming an equal-variances signal-detection model and unbiased participants with criterion $c = 0$) from the fitted true d' of each simulated participant, and sampled from binomial distributions with these probabilities the number of hits/misses and false alarms/correct rejections, for which we then calculated the observed d'. (With 96 trials, the maximum d' is 5.1233639. We truncated the true predictor values at this value, because otherwise the correlation between true and observed predictor values would be artifactually reduced.)

For the criterion, we added normally distributed measurement error with $\sigma_{Ey} = \sqrt{\sigma^2_Y - \sigma^2_{\hat{Y}}}$ (with σ^2_Y = total variance of the experimental criterion values, and $\sigma^2_{\hat{Y}}$ = variance of the simulated true criterion values). This ensured that the resulting simulated criterion values had the same mean and variance as our experimentally observed data.

5.2.1 Reliability of the Predictor and True Distributions

The above plot shows the resulting distributions of the true predictor values (orange: truncated-normal; black: gamma). Note that the truncated-normal has an additional spike at zero which is not plotted.

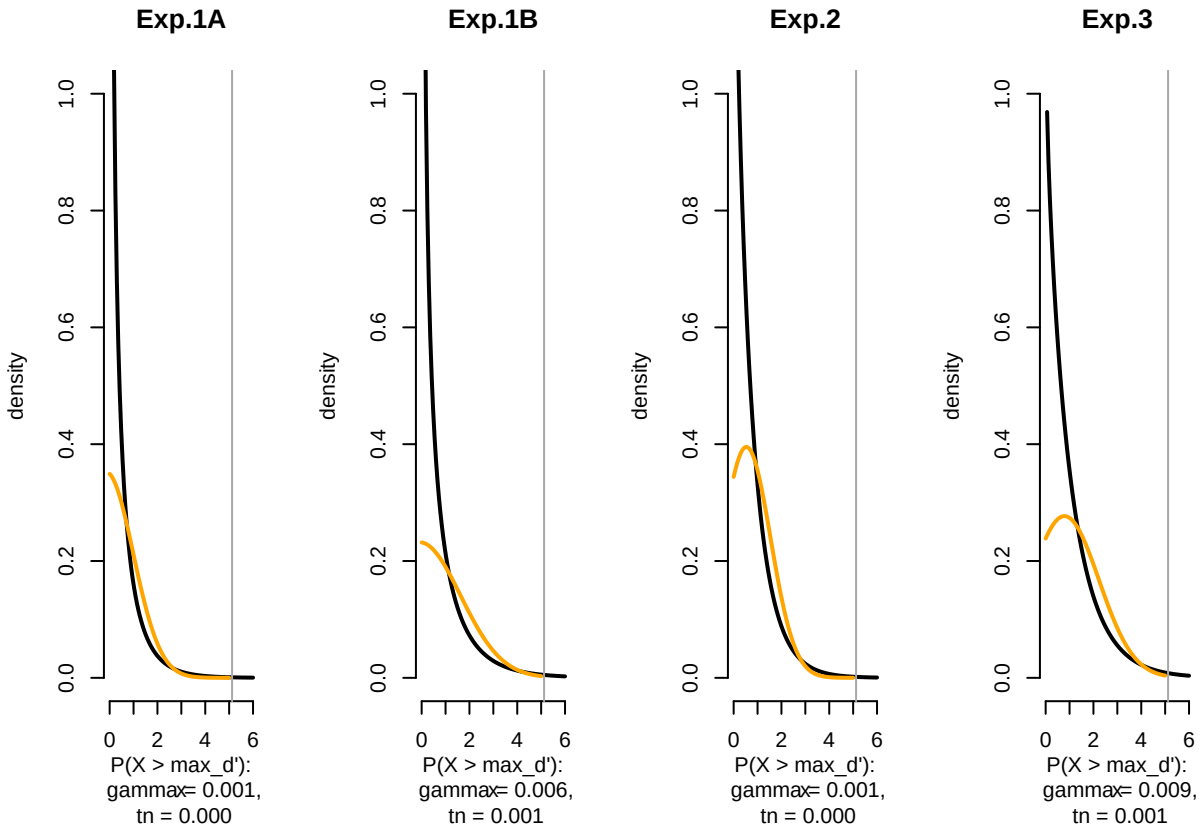

Figure 2: True predictor values with truncated normal distribution (orange) and gamma distribution (black).

```
##            variable Exp.1A Exp.1B  Exp.2  Exp.3
## 1           mu_x_nd  -0.15  -0.09   0.53   0.79
## 2        sigma_x_nd   1.13   1.72   1.01   1.44
## 3         prob_true   0.55   0.52   0.30   0.29
## 4             shape   0.39   0.44   0.91   0.94
## 5             scale   0.97   1.45   0.79   1.12
## 6          n_sample  70.00  68.00 150.00 156.00
## 7             alpha   0.00   0.00   0.00   0.00
## 8              beta   0.20   0.21   0.09   0.08
## 9          n_trials  96.00  96.00  96.00  96.00
## 10              R_XX   0.84   0.92   0.88   0.93
## 11    SD_Y_error_tn   0.18   0.05   0.18   0.14
## 12 SD_Y_error_gamma   0.18   0.06   0.18   0.14
## 13          R_YY_tn   0.33   0.94   0.13   0.25
## 14       R_YY_gamma   0.33   0.93   0.13   0.25
```

The above table shows the parameters of the predictor and criterion distributions. We fitted the truncated-normal (mu_x_nd and sigma_x_nd before truncation; mu_x and sigma_x after truncation) and gamma distributions (shape and scale) to the observed predictor values (with binomial error according to the number of trials; see predictor_fits.Rmd). The gamma distribution is strictly positive, so that the proportion true-zero (prob_true) refers only to the truncated normal distribution.

5.2.2 EiV Regression Estimation

In the EiV regression, obtaining p-values involves computations that occasionally fail, yielding missing values (specifically, the standard errors of the regression parameters could not be estimated). This problem was eliminated by using different starting values for the EiV regression algorithm.

5.3 Results

5.3.1 Fit and Descriptives

```
##       variable    dist     Exp.1A     Exp.1B      Exp.2      Exp.3
## 1  mean_xobs_m expdata 0.38188305 0.64235114 0.72350546 1.05214595
## 2  mean_xobs_m      tn 0.39287524 0.65948083 0.74179956 1.08374539
## 3  mean_xobs_m   gamma 0.39001176 0.64849422 0.74045906 1.06711163
## 4   var_xobs_m expdata 0.44642879 1.03247610 0.65629672 1.28754197
## 5   var_xobs_m      tn 0.46846151 1.08652986 0.69313329 1.37001950
## 6   var_xobs_m   gamma 0.45708133 0.94574151 0.67897866 1.21105291
## 7        rxx_m expdata 0.81814955 0.90831860 0.87138732 0.92205433
## 8        rxx_m      tn 0.83835985 0.92072983 0.88503505 0.92962066
## 9        rxx_m   gamma 0.82055342 0.90826456 0.88038273 0.92330733
## 10 mean_yobs_m expdata 0.07718978 0.13675153 0.06485526 0.07867861
## 11 mean_yobs_m      tn 0.07739503 0.13657250 0.06502086 0.07893055
## 12 mean_yobs_m   gamma 0.07691655 0.13506750 0.06520531 0.07793153
## 13  var_yobs_m expdata 0.04686103 0.04550671 0.03649778 0.02702958
## 14  var_yobs_m      tn 0.04676556 0.04504701 0.03648371 0.02684969
## 15  var_yobs_m   gamma 0.04641083 0.04108986 0.03637418 0.02630061
## 16       ryy_m expdata 0.12212338 0.07008963 0.31143161 0.17250113
## 17       ryy_m      tn 0.32695168 0.94361147 0.13181959 0.24944046
## 18       ryy_m   gamma 0.31242785 0.91392346 0.12919317 0.23021116
```

5.3.2 Intercept - Standard Regression

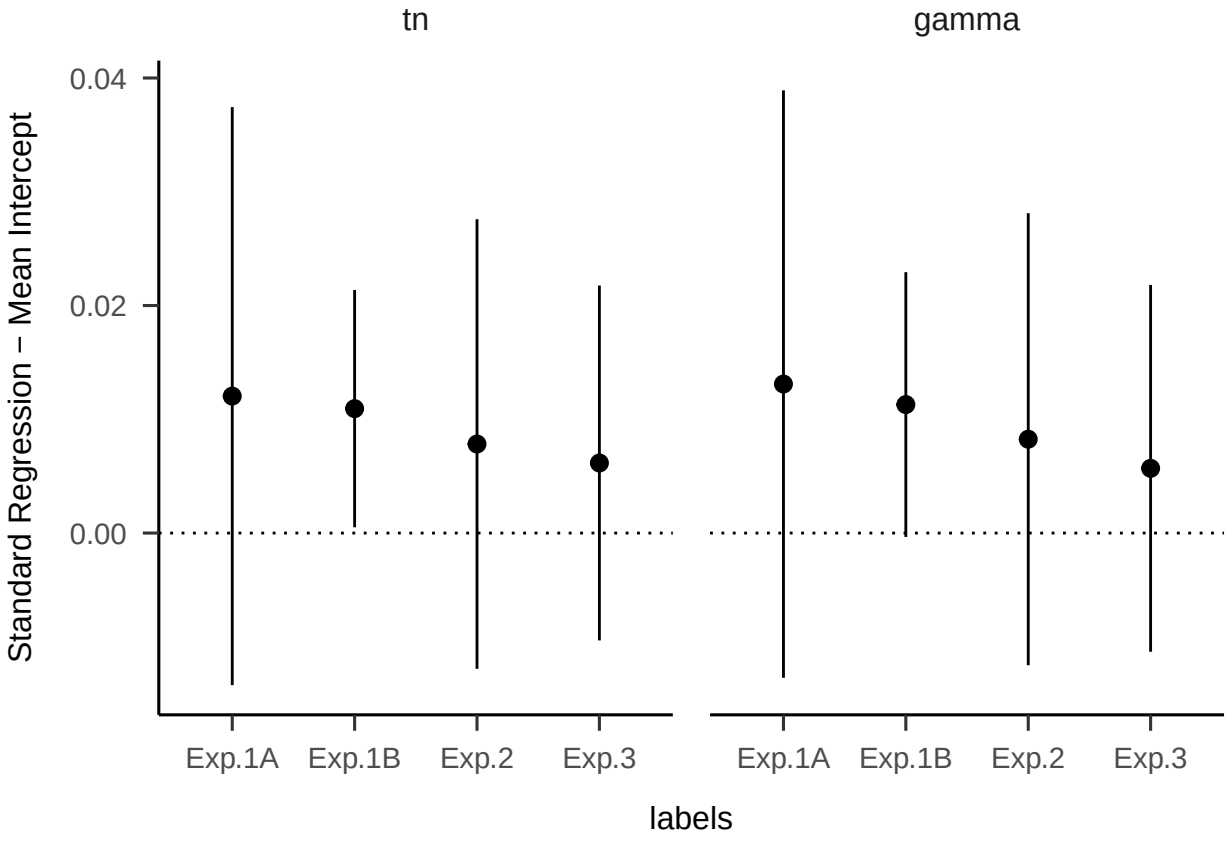

Figure 3: Intercept estimates (standard regression) as a function of experiment and distribution.

5.3.3 Intercept - EiV Correction

5.3.4 Intercept - Summary

- As expected, the intercept was overestimated in standard regression.
- The EiV correction failed to fully correct the bias for gamma-distributed predictors.

5.3.5 Alpha Inflation - One-/two-sidedness of significance tests

In the regression analyses reported in the main text of this paper, the intercept was tested two-tailed because both a significantly negative and positive intercept were of interest: A significantly positive intercept was interpreted as conditioning in the absence of CS discrimination; a significantly negative intercept would have shown that objective visibility alone is not sufficient for conditioning. Given the finding of significantly positive (but no significantly negative) intercepts, the simulations now aim at investigating whether there were more significantly positive intercepts than expected under the H_0 (not at investigating whether there are more significantly negative intercepts than expected under the H_0). This has two consequences for the simulations: First, the reported α-error rates of the simulations refer to one-tailed significantly positive, not to significantly negative intercepts (i.e., how many of the simulated intercepts are significantly positive when the true intercept is zero?). With a two-tailed test, the probability of a significantly positive (or negative) intercept is half the α-error level; in the present case, the probability is $1\%/2 = 0.5\%$ each. To keep the

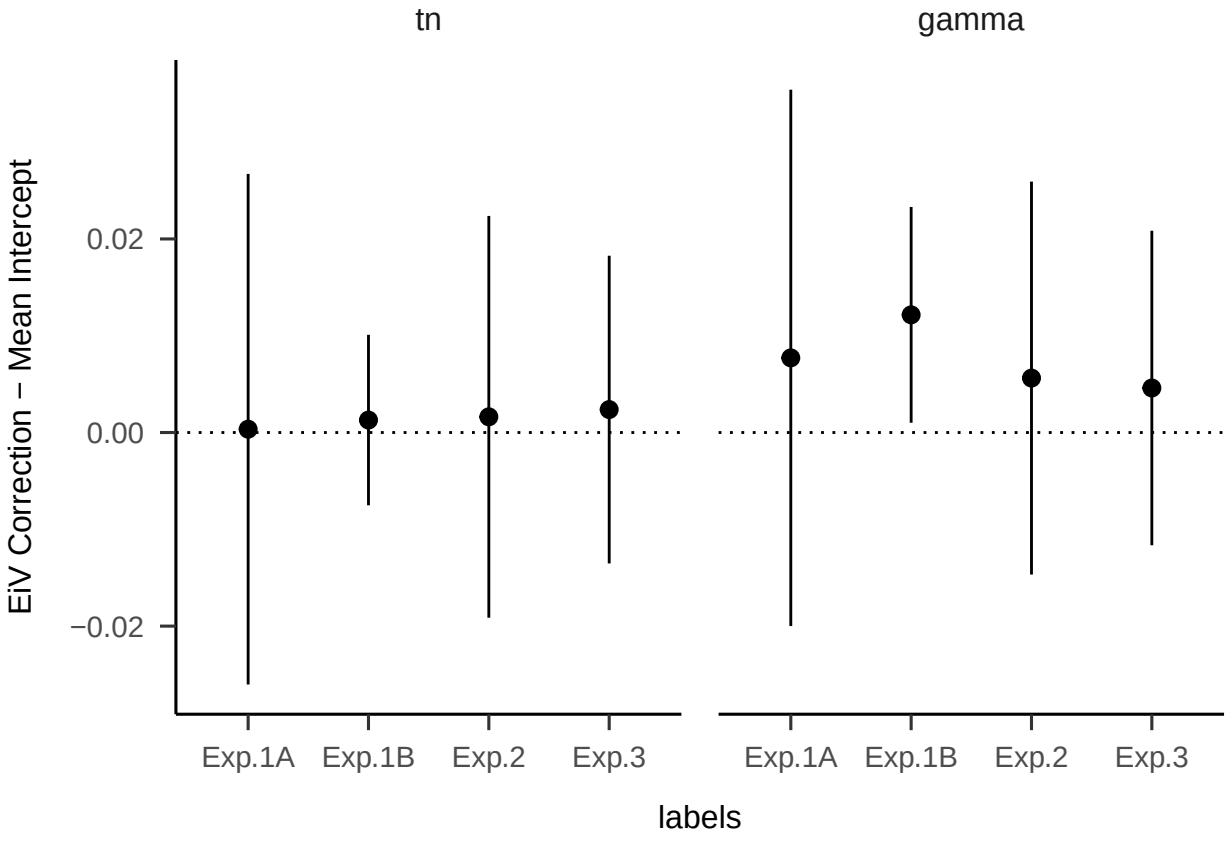

Figure 4: Intercept estimates (EiV) as a function of experiment and distribution.

probability of a significantly positive intercept identical to the previous analyses despite applying a one-tailed ($H_1 : \beta_0 > 0$) instead of a two-tailed test ($H_1 : \beta_0 \neq 0$) we set $\alpha = 0.5\%$. Second, when reporting the simulated p-value, we report the percentage of intercepts that were at least as positive as the observed intercept (and not the percentage of intercepts that were at least as extreme as the observed intercept). If this probability was less than 0.5%, the intercept was considered unlikely under the $H_0 : \beta_0 \leq 0$.

5.3.6 Alpha Inflation - Standard Regression

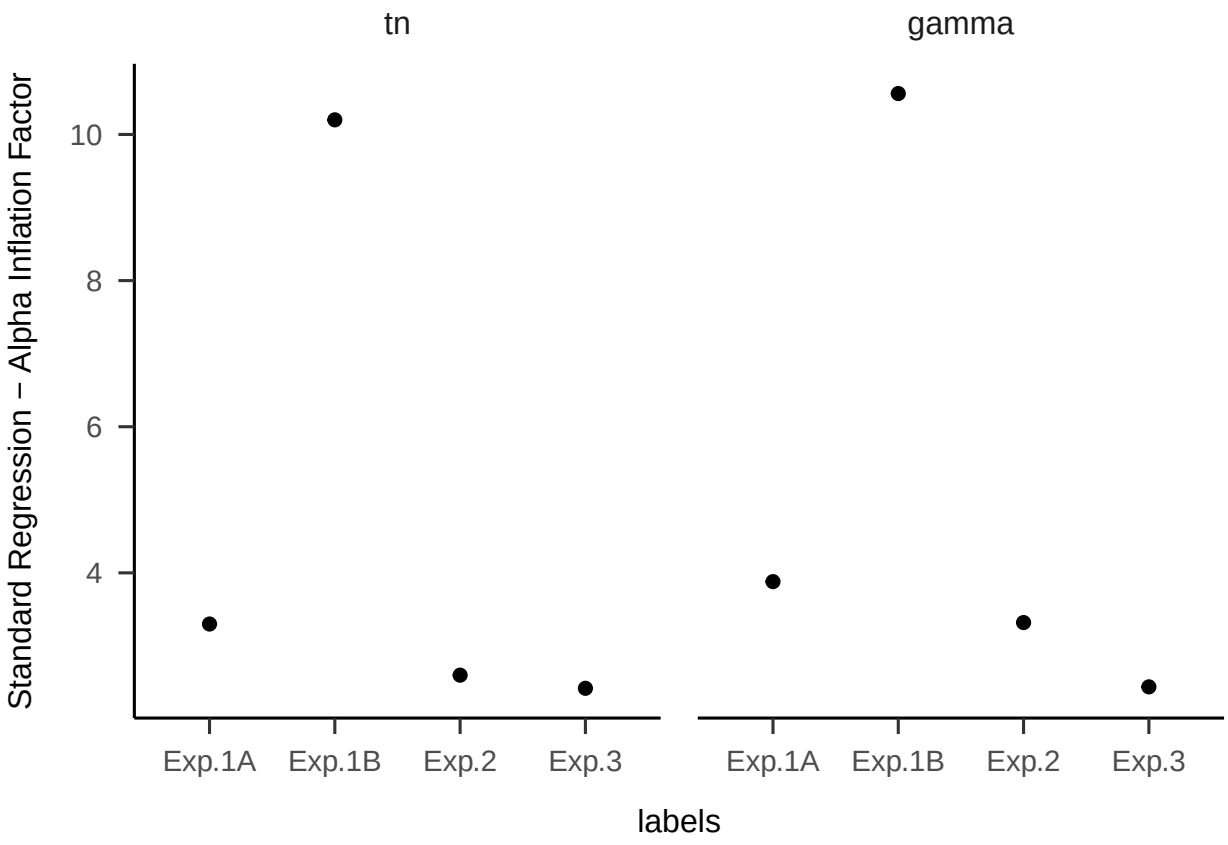

Figure 5: Inflation factors of the alpha error rate for the intercept test against zero (i.e., observed error rates divided by nominal alpha level; standard regression; for one-sided alpha = .005) as a function of experiment and distribution.

5.3.7 Alpha Inflation - EiV Correction

5.3.8 Alpha Inflation - Summary

- The alpha level was inflated for Exp. 1A and 1B in standard regression, by a factor of up to 11.
- The EiV correction nearly eliminated the inflation for a truncated-normal predictor.
- For a gamma predictor, the EiV correction reduced the inflation slightly for Exp. 1A, and even further inflated the error rate in Exp. 1B (e.g., gamma distribution, 50% unaware, high reliability of predictor and criterion), leaving a residual inflation factor of up to 18.

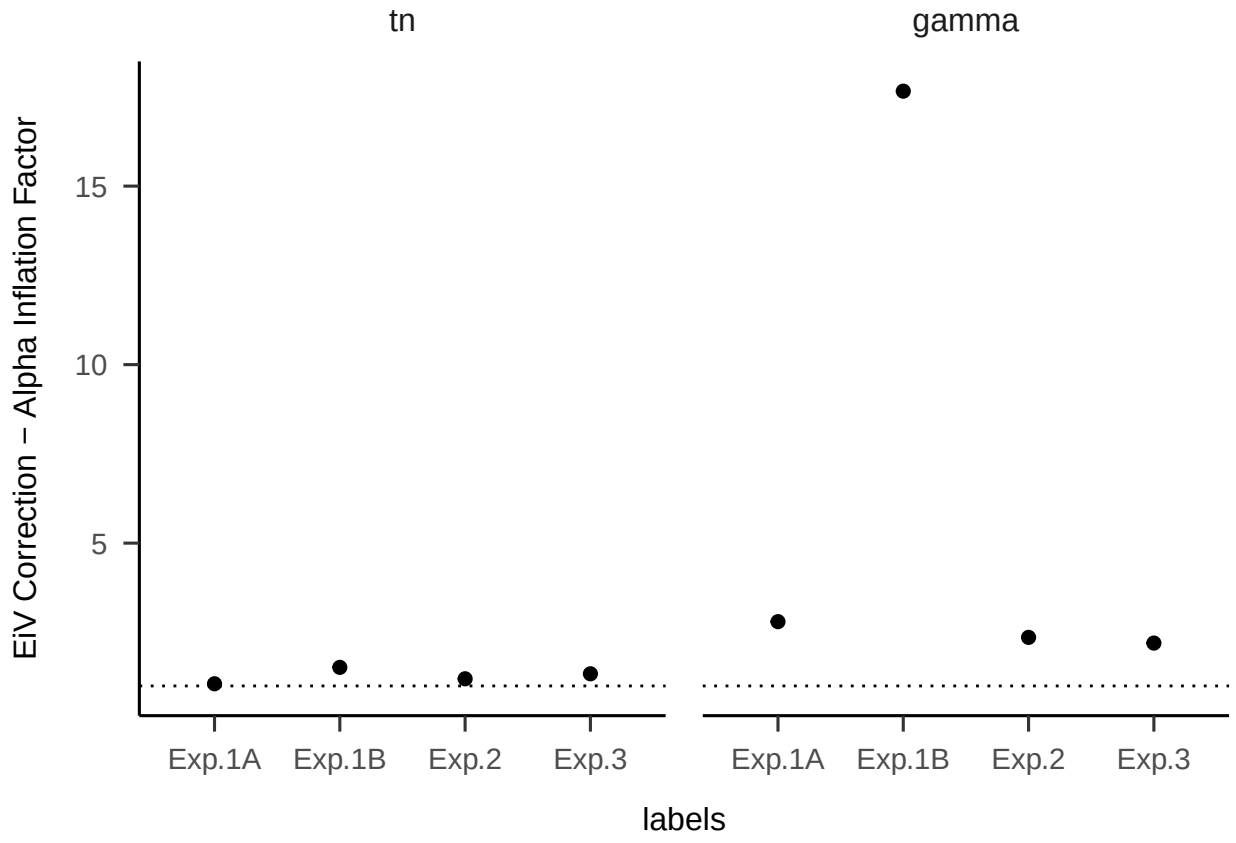

Figure 6: Inflation factors of the alpha error rate for the intercept test against zero (i.e., observed error rates divided by nominal alpha level; EiV regression; for one-sided alpha = .005) as a function of experiment and distribution.

5.3.9 Slope - Standard Regression

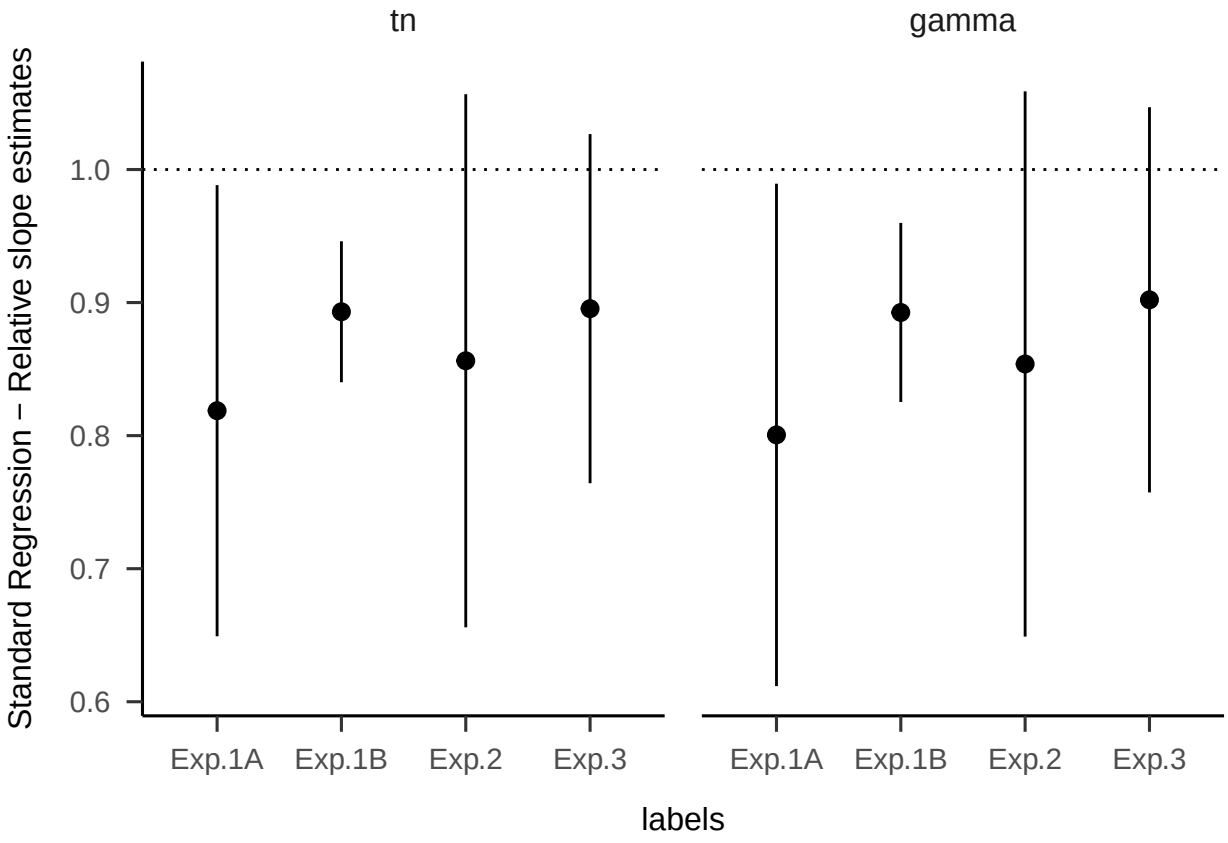

Figure 7: Relative slope estimates (i.e., mean observed slopes divided by true slope; standard regression) as a function of experiment and distribution.

5.3.10 Slope - EiV Correction

5.3.11 Slope - Summary

- The slope was underestimated by standard regression.
- The EiV correction generally worked quite well, but failed to fully correct for this bias.

5.4 Slope Power - Standard Regression

5.4.1 Slope Power - EiV Correction

5.4.2 Slope Power - Summary

- The power of the slope test was very high.

5.4.3 Simulated p-Values

```
##   labels  dist std_intercept_Psim eiv_intercept_Psim
```

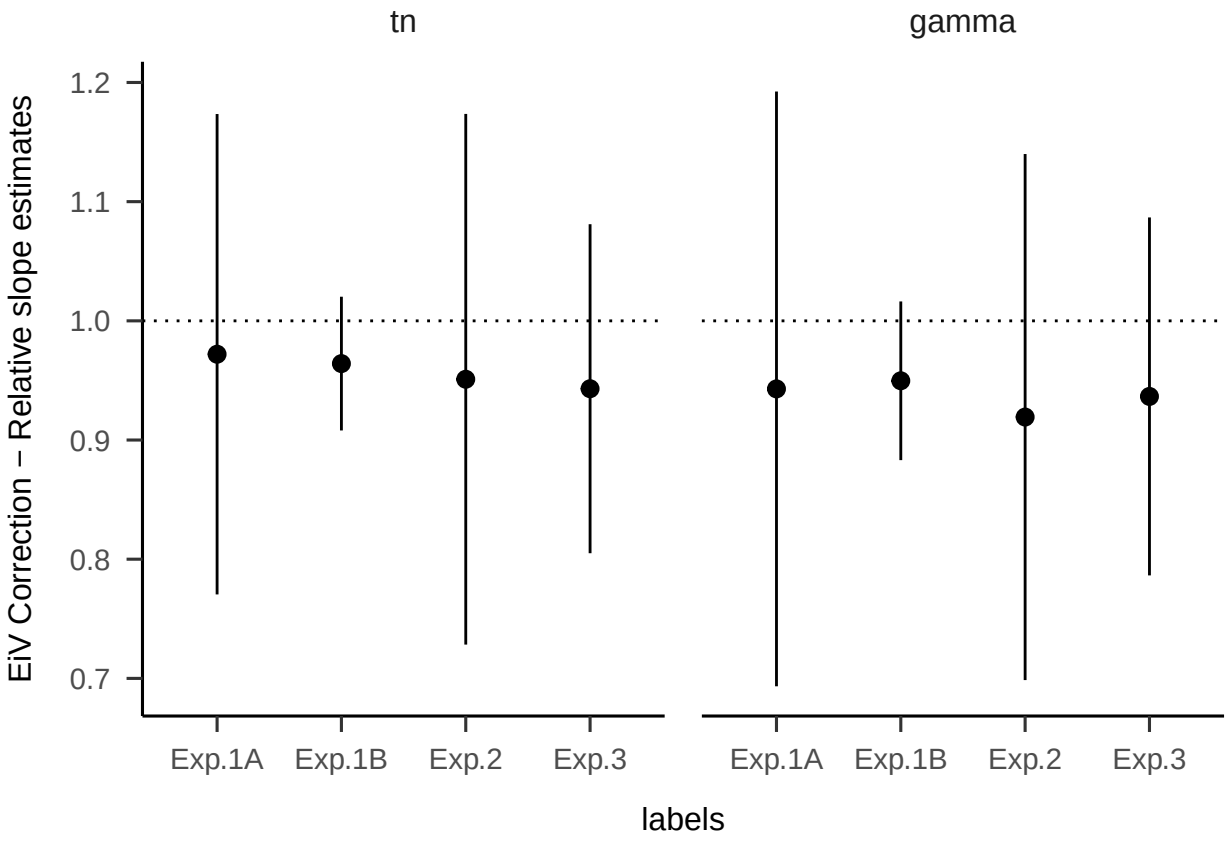

Figure 8: Relative slope estimates (i.e., mean observed slopes divided by true slope; EiV regression) as a function of experiment and distribution.

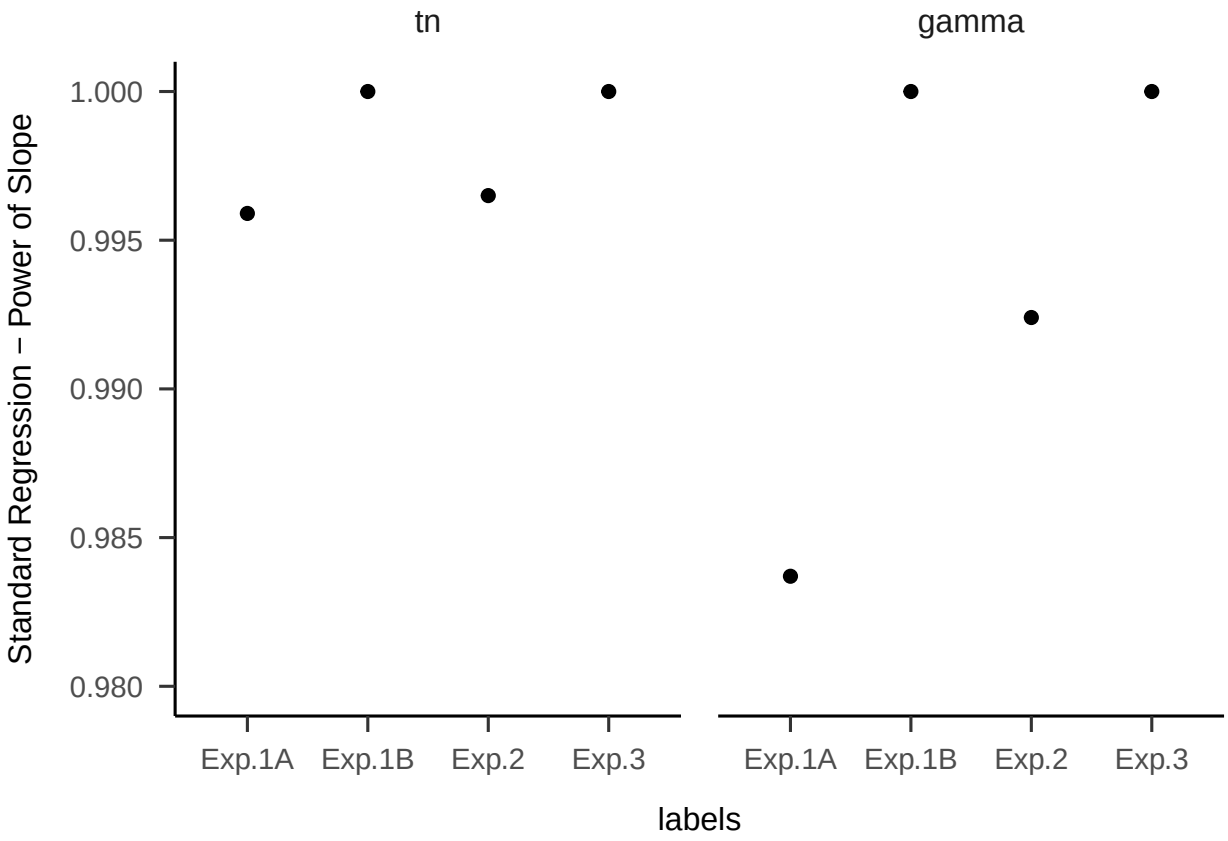

Figure 9: Power of the intercept test (i.e., proportion of results significant at alpha=.05; standard regression) as a function of experiment and distribution.

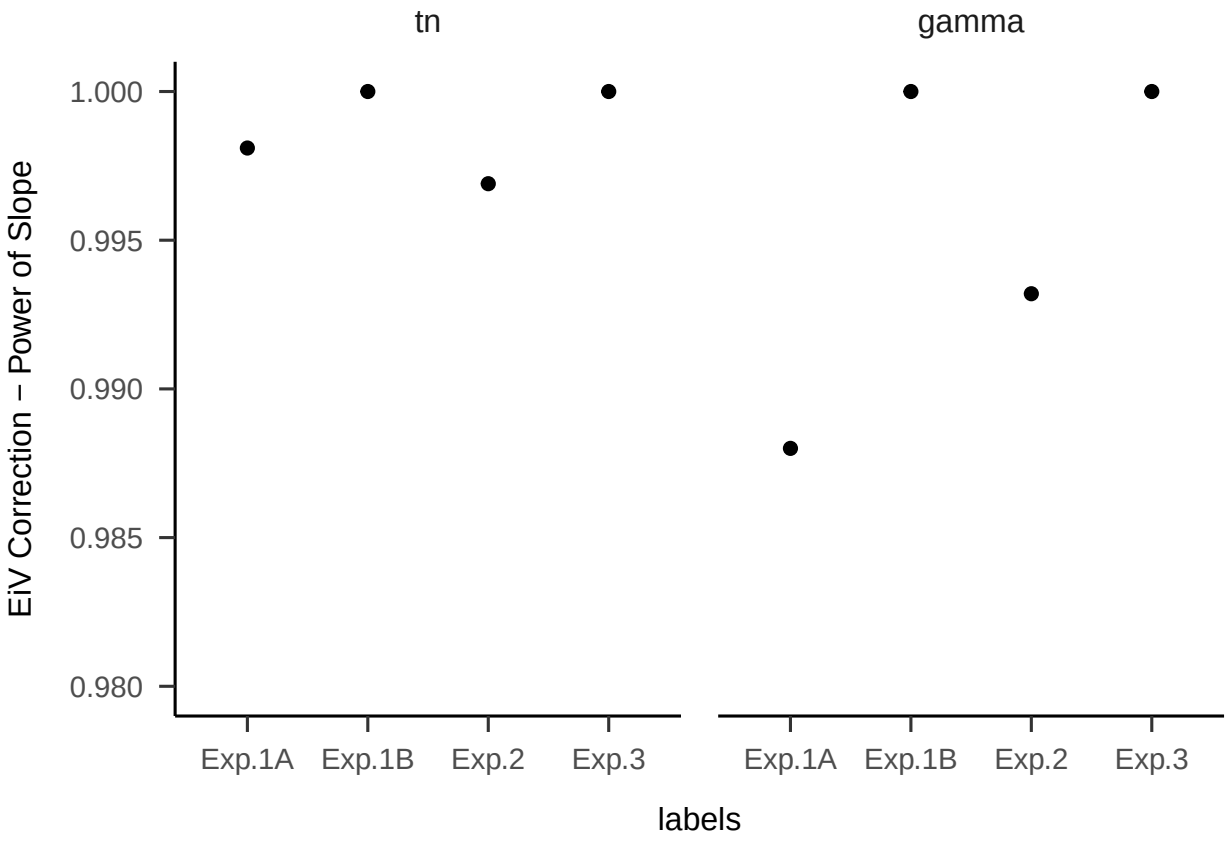

Figure 10: Power of the intercept test (i.e., proportion of results significant at alpha=.05; EiV regression) as a function of experiment and distribution.

```
## 1     1A gamma               .052               .037
## 2     1A    tn               .046               .020
## 3     1B gamma             < .001             < .001
## 4     1B    tn             < .001             < .001
## 5      2 gamma               .018               .014
## 6      2    tn               .015               .008
## 7      3 gamma             < .001             < .001
## 8      3    tn             < .001             < .001
##   std_intercept_P_onesided_exp eiv_intercept_P_onesided_exp
## 1                         .034                         .033
## 2                         .034                         .033
## 3                       < .001                       < .001
## 4                       < .001                       < .001
## 5                         .009                         .009
## 6                         .009                         .009
## 7                       < .001                       < .001
## 8                       < .001                       < .001
```

The experimental intercept was significant with standard regression and EiV correction for Exp. 1A and Exp. 3, but not for Exp. 1B and Exp. 2 (with $\alpha_{onesided} = .005$). The same intercepts were (non-)significant when comparing the experimental intercepts with the distributions of simulated intercepts.

6 EiV Robustness Checks, Part 2 (SDT Model)

6.1 Rationale

In a first set of simulations with data similar to those we observed, we found that the EiV correction worked well. This finding is consistent with the successful validation results reported by Klauer et al., but inconsistent with those of Miller. Here we simulated a broader set of conditions and aimed at identifying the factors that affect EiV performance, and at identifying scenarios/conditions under which it may fail.

6.2 Model

The general approach was similar to the first set of simulations: We assumed a single-process null model with a zero intercept and a perfect correlation between predictor (CS discrimination) and criterion (learning effect).

We varied the predictor's distributional form as well as its parameters: True predictor values were drawn either from a truncated-normal (as assumed by Klauer's EiV correction) or a gamma distribution (which fitted the present data well but violates the distributional assumptions underlying the Klauer EiV correction). The mean of the normal (before truncation) was varied in three steps, resulting in three conditions with different proportions of true-zero predictor values (.1, .5, .9). They were paralleled by three gamma distributions whose means and variances were identical to those of the truncated-normal distributions.

We varied the reliability of the predictor (CS discrimination) via the number of trials used to assess CS discrimination (40, 80, 160, 320). We computed the observed predictor values based on an equal-variance signal-detection model without response bias (criterion = 0): True predictor d' values were converted into true π(Hit) and π(FA), and hits and false alarms were drawn from a binomial distribution. The resulting observed p(Hit) and p(FA) were then used to compute the observed d' predictor values. (Note that the resolution of this model depends on the number of trials. It produces artifacts with smaller trial numbers when the predictor can assume high values with nonnegligible proportions. We selected appropriate simulation parameters that minimized the occurrence of this issue, and addressed it where it could not be avoided by truncating the remaining true predictor values at the maximum d', to preclude a biasing of the error variance. With the minimum of 40 trials selected here, the maximal d' was 4.48, and the proportion of affected cases was <1% for the gamma distribution in the 10% true-zero proportion case and <0.1% in all other cases.)

The reliability of the criterion (conditioning effect) was varied in three steps (.1, .3, .5), covering a range representative of masked-priming effects (https://doi.org/10.5334/joc.419). Normally distributed measurement error was added to the true criterion values (its standard deviation was computed on the basis of the reliability and the true criterion variance).

Finally, the regression slope was varied in three steps (0.25, 0.5, 1), ranging from a relatively low value (similar to that observed in our data) to higher values (more representative of Miller's simulations). Note that the model ensures a perfect latent correlation between predictor and criterion, so the slope does not represent the strength of the relation but merely a scaling factor (i.e., the ratio of the means, mu_y/mu_x, and of the standard deviations, sigma_y/sigma_x).

All parameters refer to the population, and the sample characteristics differed from population parameters due to sampling error.

In sum, the manipulated parameters were:

- predictor distribution type: truncated-normal or gamma (tn vs. gamma)
- predictor proportion true-zero (.1, .5, .9)
- reliability of the predictor (via trial number: 40, 80, 160, 320)
- reliability of the criterion (.1, .3, .5)
- slope (0.25, 0.5, 1)

- sample size (50, 150, 500)

Constants:

- The variance of the normal distribution before truncation was constant (sigma_x_nd=1).
- The regression intercept was fixed to zero.
- In each cell of the simulation, there were n_sim=1000 samples.

6.2.1 True Predictor Distributions

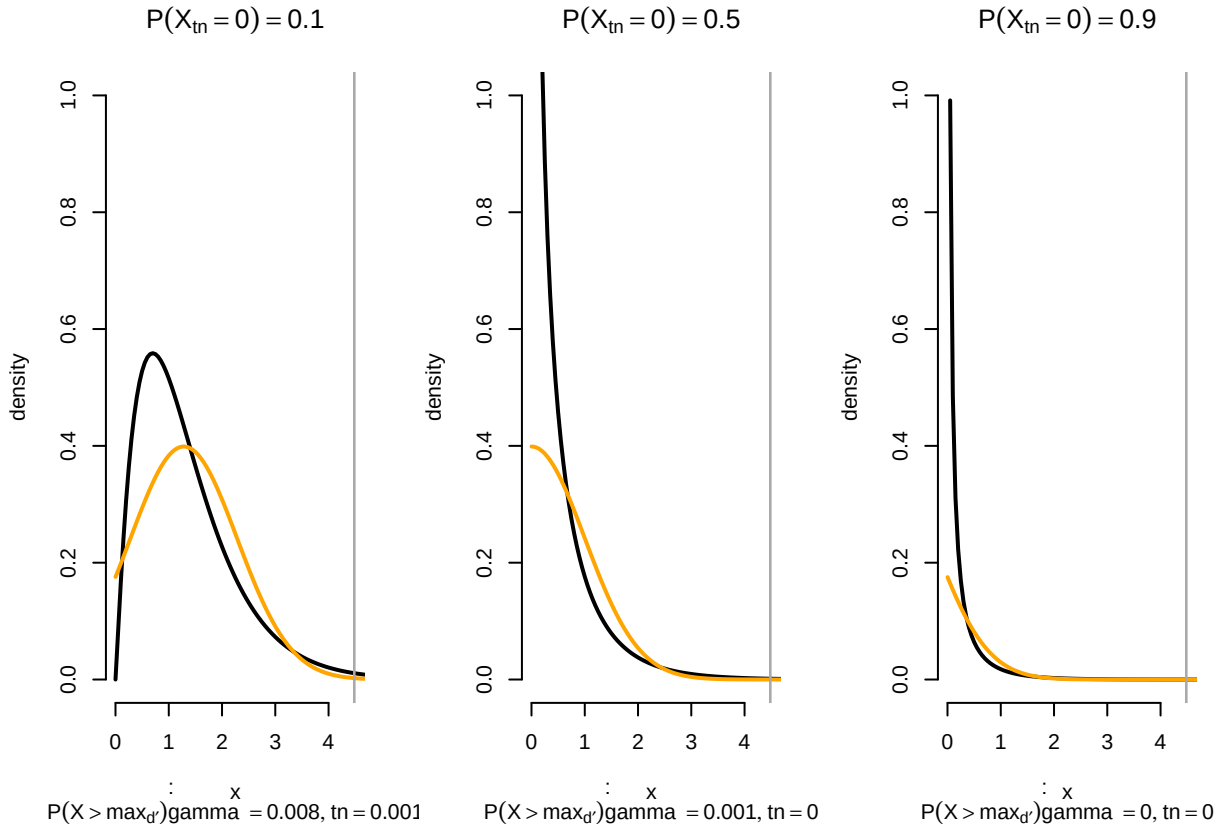

Figure 11: True predictor values with truncated normal distribution (orange) and gamma distribution (black).

For each of the two distribution types, three parameter sets were realized. The above plot shows the resulting distributions of the true predictor values (orange: truncated-normal; black: gamma). Note that the truncated-normal has an additional spike at zero that is not plotted.

```
##   variable Proportion true zero: 0.1 Proportion true zero: 0.5
## 1  mu_x_nd                      1.28                      0.00
## 2     mu_x                      1.33                      0.40
## 3  sigma_x                      0.91                      0.58
## 4    shape                      2.11                      0.47
## 5    scale                      0.63                      0.85
##   Proportion true zero: 0.9
```

```
## 1                    -1.28
## 2                     0.05
## 3                     0.19
## 4                     0.06
## 5                     0.78
```

The standard deviation of the normal distribution before truncation (sigma_x_nd) was 1 for all three plotted distributions, and the means were set so that the proportion of true zero predictor values was .1, .5 and .9. The resulting standard deviation of the true predictor values after truncation (sigma_x) varied strongly between the three distributions (between $\sigma_x = 0.19$ for 90% true zero and $\sigma_x = 0.91$ for 10% true zero). The shape and scale parameters of the gamma distribution were selected such that the resulting mu_x and sigma_x were identical to those of the respective truncated-normal. Note that the gamma distribution is strictly positive; the 'true-zero proportion' property refers only to the truncated-normal distributions.

6.2.2 Reliability of the Predictor

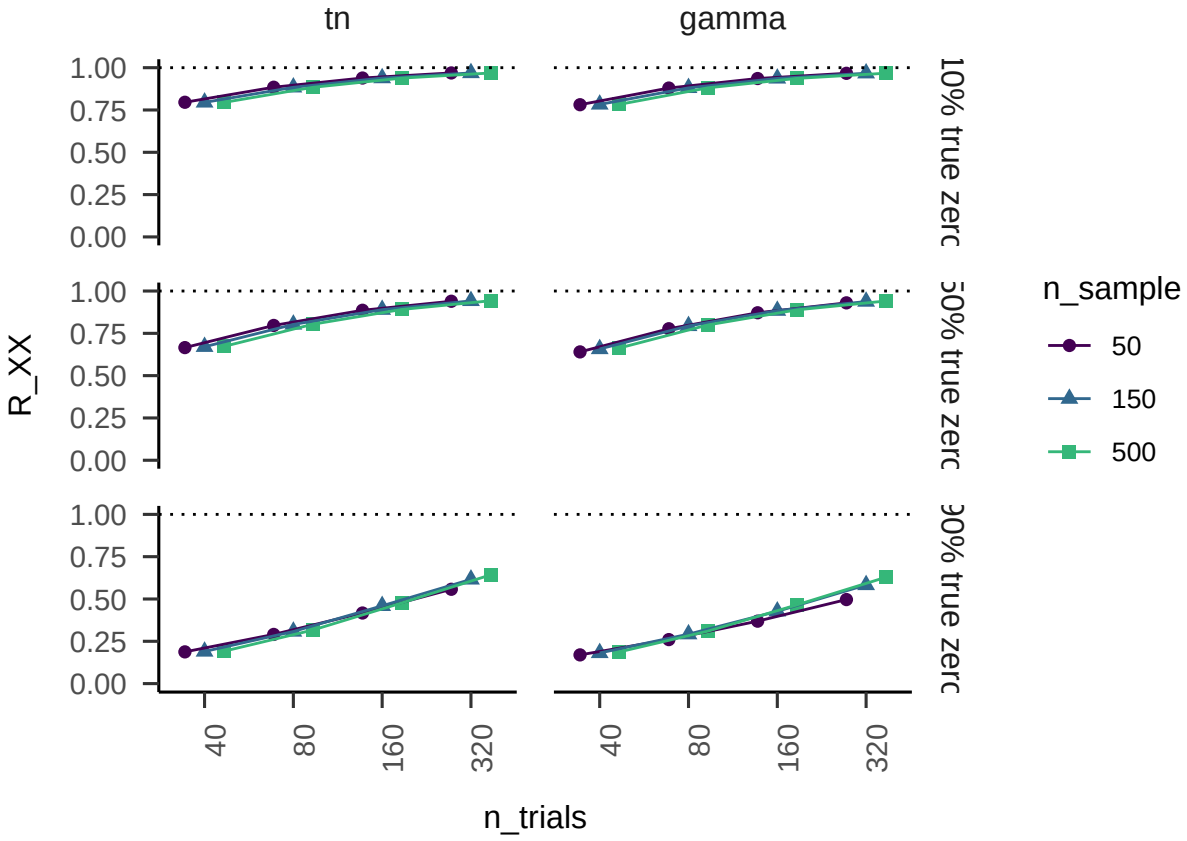

Figure 12: Reliability of predictor as a function of measurement error (n_trials) and sample size (n_sample), split by distribution type (columns), and true-zero proportion (rows).

The reliability of the predictor was manipulated by varying the number of trials. Reliability was estimated by sampling observed predictor values and computing their squared correlation with the true predictor values. The resulting reliability estimates were affected by the number of trials and the proportion of true-zero predictor values. They are plotted here (as a function of distribution type, proportion true-zero, n_trials, n_sample), and range from very low (for 90% true zero and few trials) to very high (for 10% true zero and many trials; distribution type and sample size had little to no effect).

6.2.3 EiV Regression Estimation

In the EiV regression, obtaining p-values involves computations that occasionally fail, yielding missing values (specifically, the standard errors of the regression parameters could not be estimated). This problem was reduced but not eliminated by using different starting values for the EiV regression algorithm. If the algorithm did not succeed after trying out 10 different starting values, a new sample was drawn and EiV regression was again attempted; and this procedure was repeated until a sample was obtained for which EiV regression returned the standard error estimates necessary for computing p-values.

This replacement procedure did not affect the results. We ran a simulation with missing values, and compared the results to those of a simulation in which missing values were replaced. The results did not differ meaningfully: Across all dependent variables, all the predictors involving the replacement factor had very small effects (ges<.001).

6.3 Main Results

6.3.1 Intercept - Standard Regression

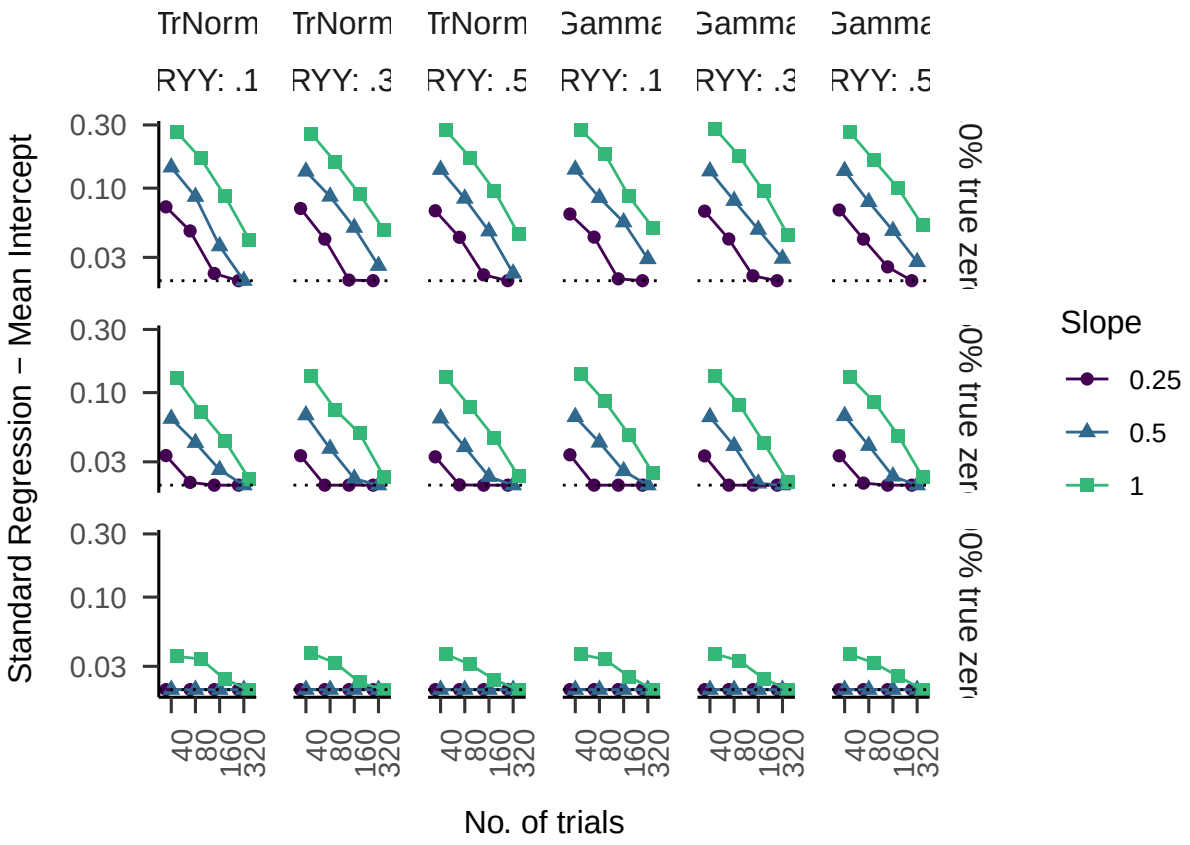

Figure 13: Intercept estimates (standard regression; for N = 150) as a function of measurement error (n_trials) and slope, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

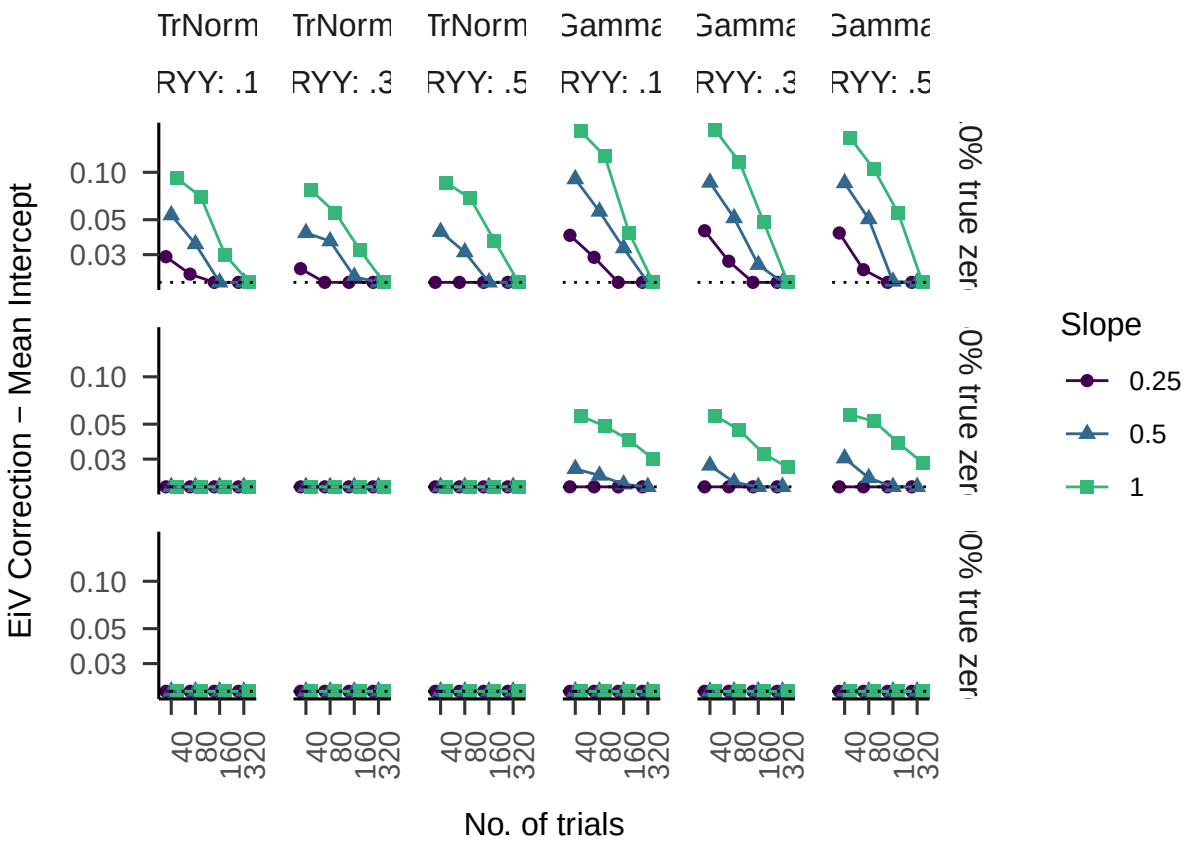

Figure 14: Intercept estimates (EiV regression; for N = 150) as a function of measurement error (n_trials) and slope, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

6.3.2 Intercept - EiV Correction

6.3.3 Intercept - Summary

- The intercept was overestimated in standard regression in the presence of measurement error.
- The magnitude of this bias depended on: proportion true-zero (prob_true), measurement error (n_trials), slope (beta)
- The EIV correction failed to fully correct the bias, leaving residual bias.
- The magnitude of this residual bias depended on: distribution type (dist), proportion true-zero (prob_true), measurement error (n_trials), slope (beta). (There was also a small effect of n_sample (not shown), with smaller residual bias for smaller samples and stronger residual bias for larger samples.)
- The EiV correction worked better for truncated-normal and less well for gamma-distributed predictors (even in the 50% unaware case).
- But even for a truncated-normal predictor, the EiV method failed to fully correct the intercept in cases with (medium or) small proportion true-zero.
- This residual bias was strongly reduced, but not eliminated, in another set of simulations that used a normally distributed error of the predictor (instead of the SDT model used here).
- Thus, the residual bias can be (mostly) explained by the violation of the EiV's assumption of a normally distributed error.

6.3.4 Alpha Inflation - Standard Regression

6.3.5 Alpha Inflation - EiV Correction

6.3.6 Alpha Inflation - Summary

- The alpha level was inflated in many cases in standard regression, by a factor of up to 94.
- The magnitude of inflation depended on: proportion true-zero (prob_true), measurement error of predictor (n_trials) and criterion (R_YY), sample size (n_sample)
- The EiV correction reduced but failed to fully eliminate the inflation, leaving a residual inflation factor of up to 45.
- The magnitude of the residual bias depended on: distribution type (dist), measurement error of predictor (n_trials)
- Note that the EiV correction may even further inflate the error rate in certain cases (e.g., gamma distribution, 50% unaware, high reliability of X and Y).
- The EiV correction generally worked well for a truncated-normal predictor; however, up to 4.20-fold inflation remained in the 50%-unaware case, and up to approx. 14-fold alpha inflation remained in the 10%-unaware case when criterion reliability was higher.
- This residual inflation is largely due to the violation of the normal-error assumption. In a simulation with normally distributed error, alpha inflation of EiV regression results was reduced to a factor of approximately 2-3.
- The EiV correction performed much worse for a gamma-distributed predictor, especially for larger samples and a more reliable criterion; the remaining inflation had a factor of up to 7.10 (for low criterion reliability) and increased to 22.40 (45.00) for medium (high) criterion reliability.

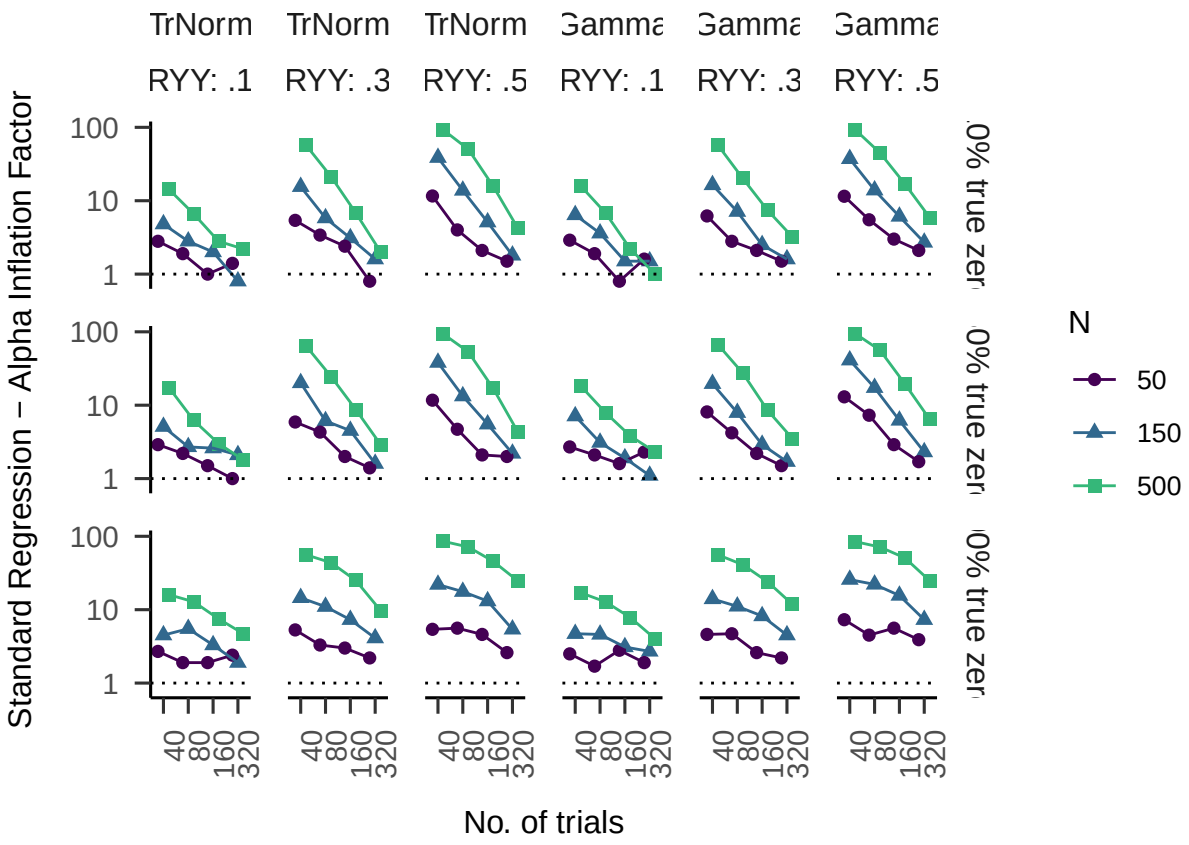

Figure 15: Inflation factors of the alpha error rate for the intercept test against zero (i.e., observed error rates divided by nominal alpha level; standard regression; for one-sided alpha = .01 and slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

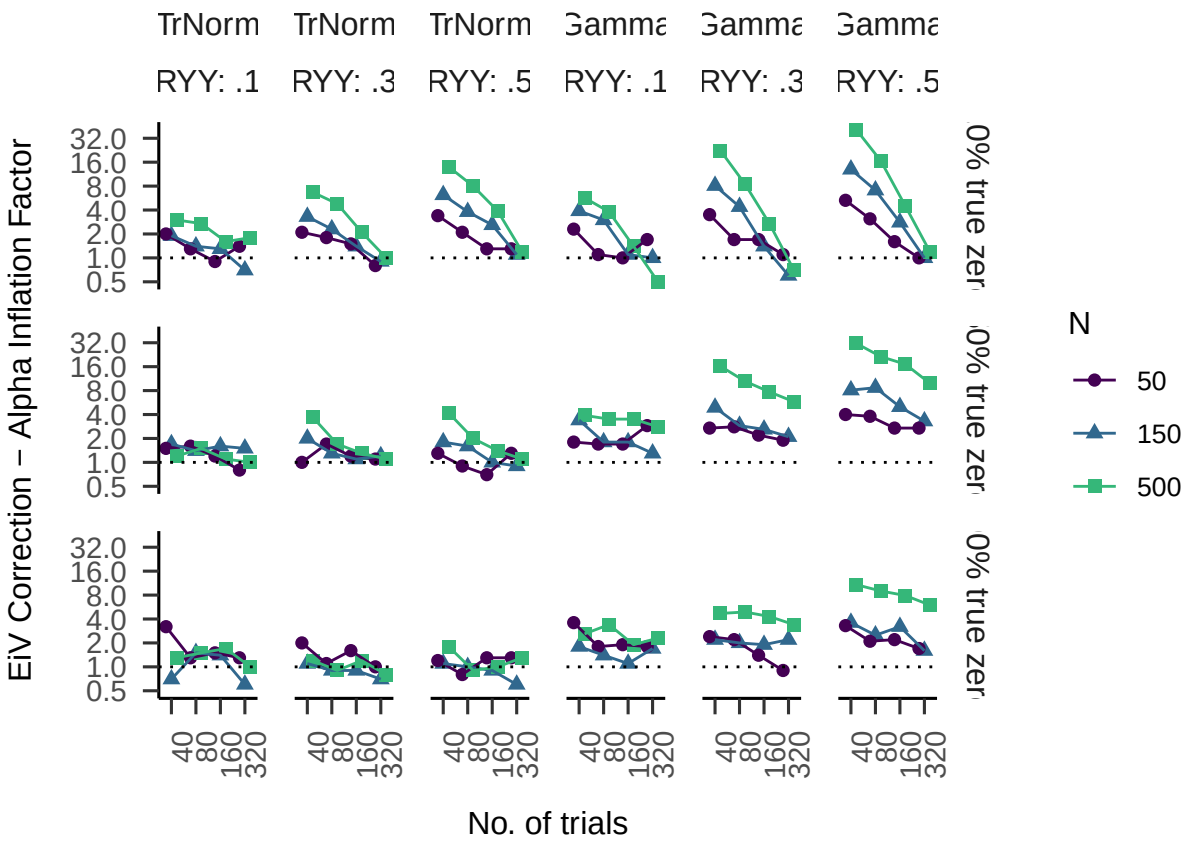

Figure 16: Inflation factors of the alpha error rate for the intercept test against zero (i.e., observed error rates divided by nominal alpha level; EiV regression; for one-sided alpha = .01 and slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

6.3.7 Slope - Standard Regression

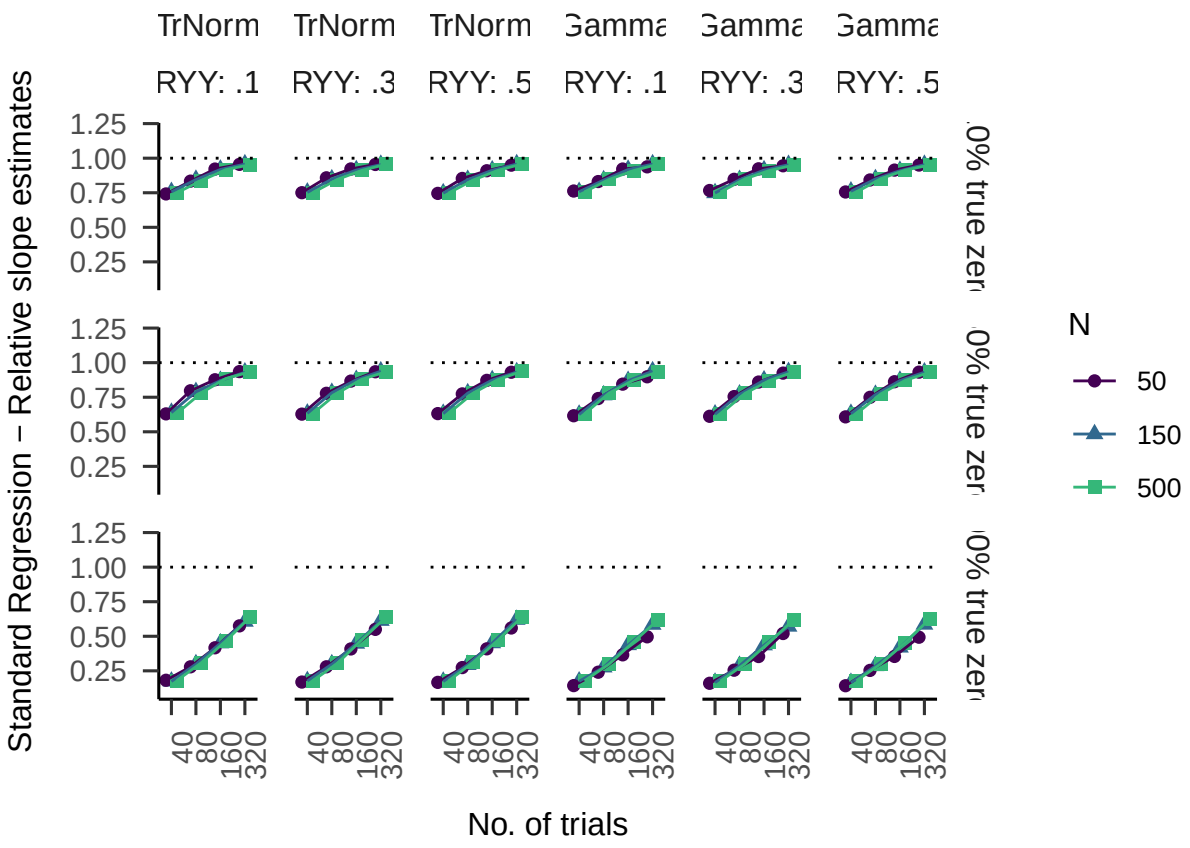

Figure 17: Relative slope estimates (i.e., mean observed slopes divided by true slope; standard regression; for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

6.3.8 Slope - EiV Correction

6.3.9 Slope - Summary

- In the presence of measurement error in the predictor, the slope was underestimated by standard regression.
- Computed as a proportion of the true slope, the magnitude of this bias depended on: proportion true-zero (prob_true), measurement error in the predictor (n_trials), sample size (n_sample)
- The EiV correction generally worked quite well, but failed to fully correct for this bias; a small residual bias remained in many cases.
- The magnitude of the residual bias depended on all factors: distribution type (dist), proportion true-zero (prob_true), measurement error in the predictor (n_trials) and in the criterion (R_YY), sample size (n_sample), slope (beta) todo: auch mit normal error noch bias übrig?
- The residual bias was larger for cases with smaller values of proportion true-zero and larger measurement error (i.e., smaller trial numbers); and it was somewhat larger for a gamma-distributed predictor, especially in cases with lower proportion true-zero and smaller trial numbers.

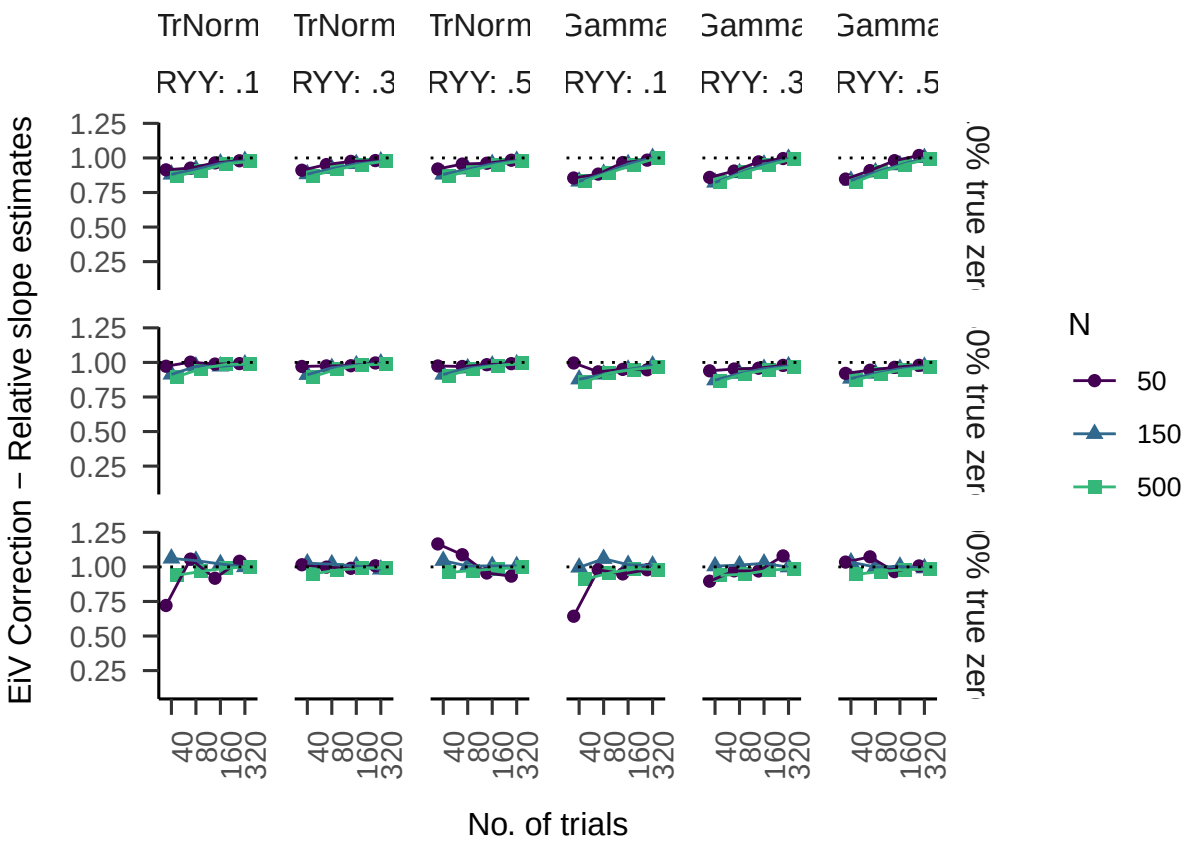

Figure 18: Relative slope estimates (i.e., mean observed slopes divided by true slope; EiV regression; for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

6.3.10 Slope Power - Standard Regression

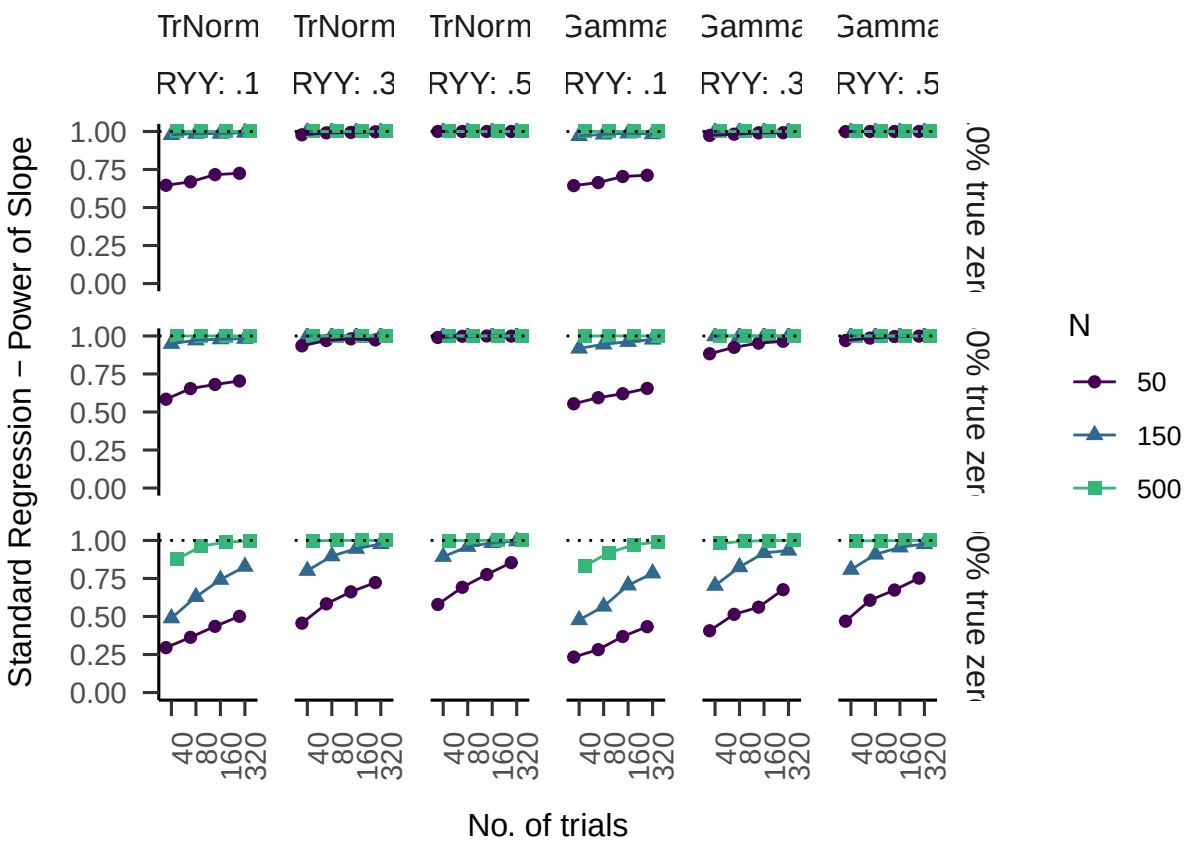

Figure 19: Power of the intercept test (i.e., proportion of results significant at one-sided alpha=.05; standard regression, for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

6.3.11 Slope Power - EiV Correction

6.3.12 Slope Power - Summary

- The power of the slope test depended mainly on sample size, true-zero proportion, and the reliability of predictor and criterion.
- The EiV correction brought a small increase in power (especially for the 90%-true-zero scenario, which was the scenario with the greatest potential for improvement).
- While power was high in many cases, it was below acceptable in several relevant scenarios that resembled G&DH's data (i.e., high proportion true-zero, low criterion reliability, smaller samples). In these cases, even perfect slopes will regularly be missed.

6.4 Additional Results

6.4.1 Alpha Inflation - EiV Correction: Alpha Level

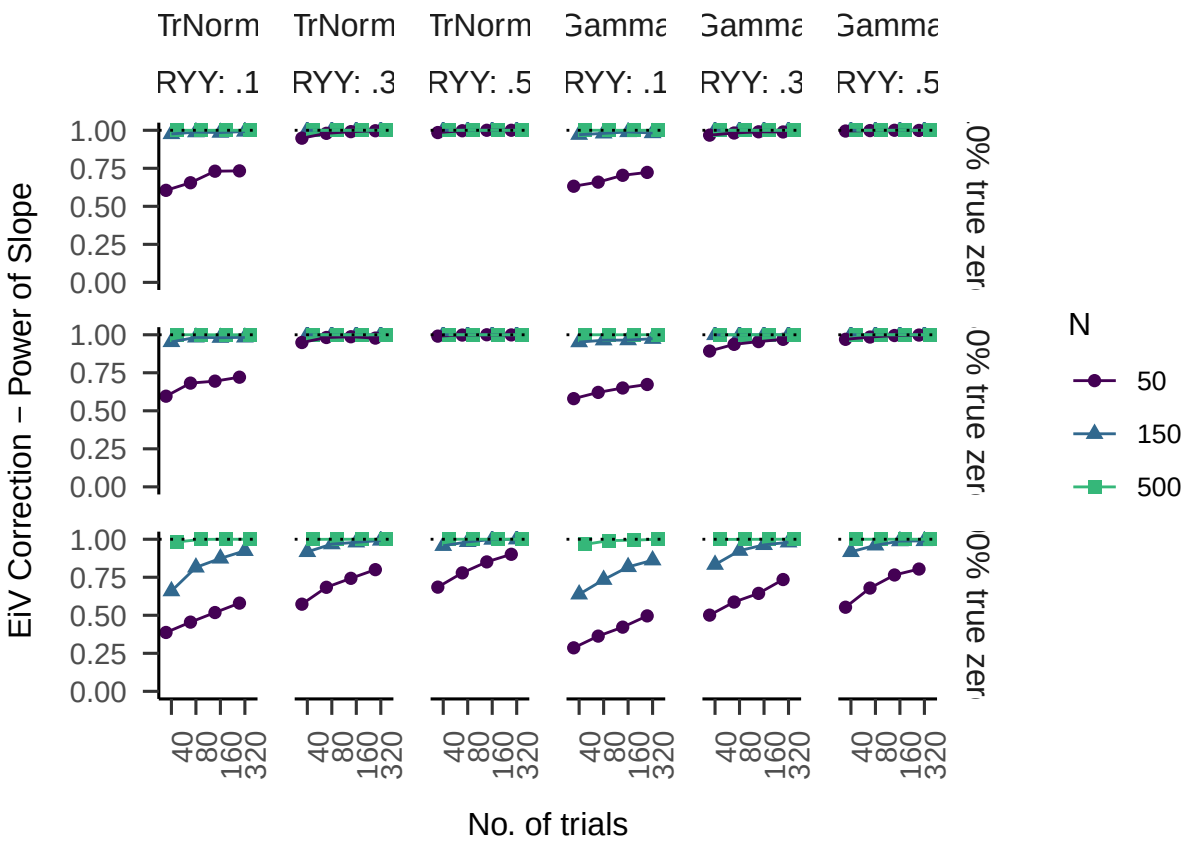

Figure 20: Power of the intercept test (i.e., proportion of results significant at one-sided alpha=.05; EiV regression; for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

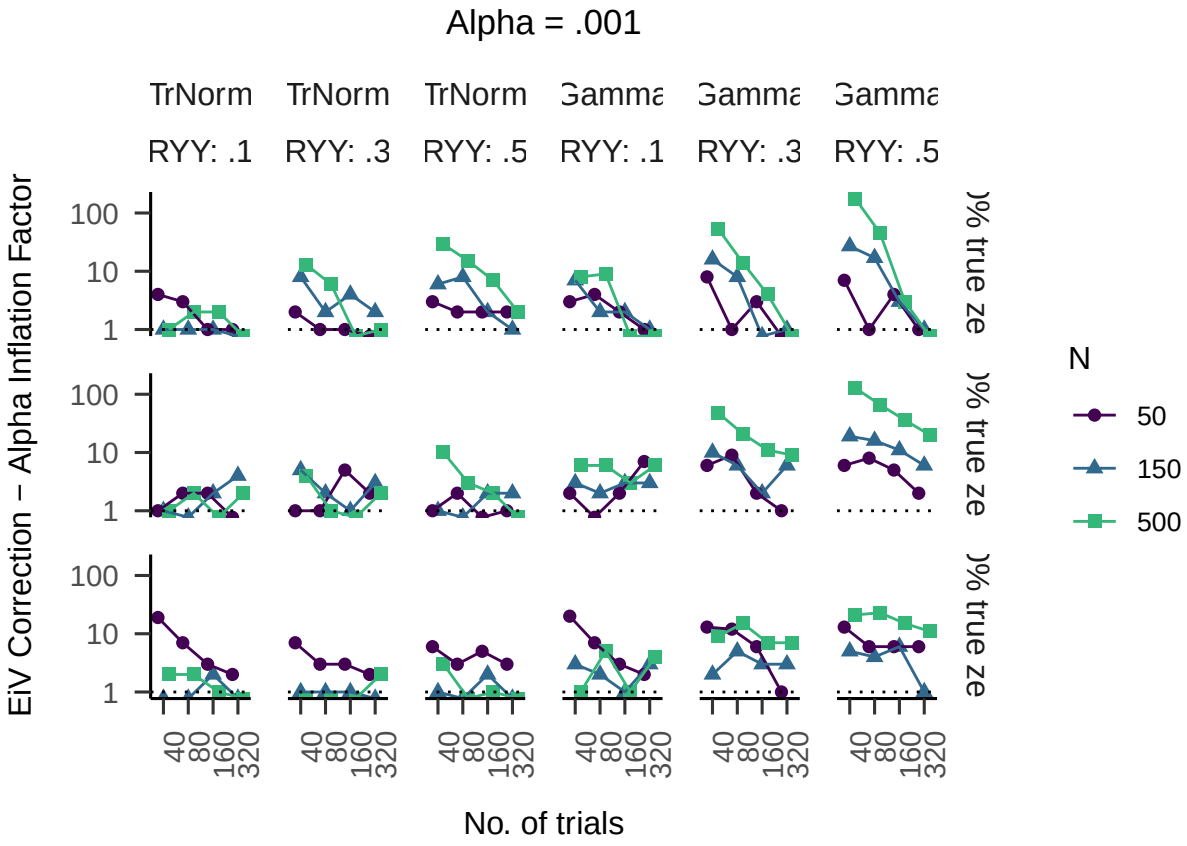

Figure 21: Inflation factor increased with decreasing alpha level.

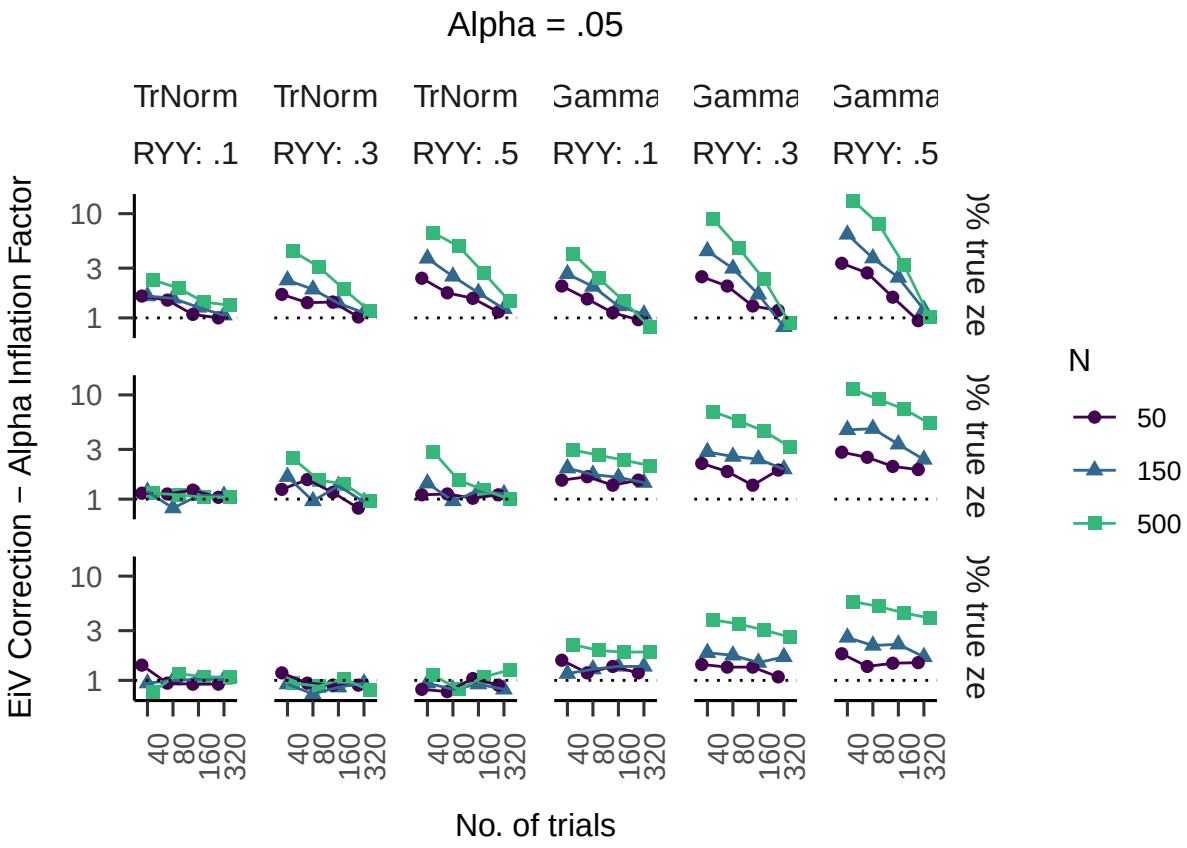

Figure 22: Inflation factor increased with decreasing alpha level.

7 EiV Robustness Checks, Part 2 (Normal Measurement Error)

7.1 Rationale

In a previous simulation with observed values generated via the SDT model, we found a residual bias when applying the EiV correction. In this set of simulations we aimed to test whether this residual bias is eliminated when directly sampling normally distributed error.

7.2 Model

The general approach was similar to the first set of simulations: We assumed a single-process null model with a zero intercept and a perfect correlation between predictor (CS discrimination) and criterion (learning effect).

We varied the predictor's distributional form as well as its parameters: True predictor values were drawn either from a truncated-normal (as assumed by Klauer's EiV correction) or a gamma distribution (which fitted the present data well but violates the distributional assumptions underlying the Klauer EiV correction). The mean of the normal (before truncation) was varied in three steps, resulting in three conditions with different proportions of true-zero predictor values (.1, .5, .9). They were paralleled by three gamma distributions whose means and variances were identical to those of the truncated-normal distributions.

We varied the reliability of the predictor (CS discrimination measure) via the number of trials used to assess CS discrimination (40, 80, 160, 320). For comparability with the SDT-model simulations, the magnitude of error variance was estimated based on the SDT model, but observed predictor values were computed by adding normal error (with SD estimated from the SDT model) to the true predictor values.

The reliability of the criterion (learning effect) was varied in three steps (.1, .3, .5), covering a range representative of masked-priming effects (https://doi.org/10.5334/joc.419). Normally distributed measurement error was added to the true criterion values (its standard deviation was computed on the basis of the reliability and the true criterion variance).

Finally, the regression slope was varied in three steps (0.25, 0.5, 1), ranging from a relatively low value (similar to that observed in our data) to higher values (more representative of Miller's simulations). Note that the model ensures a perfect latent correlation between predictor and criterion, so the slope does not represent the strength of the relation but merely a scaling factor (i.e., the ratio of the means, mu_y/mu_x, and of the standard deviations, sigma_y/sigma_x).

All parameters refer to the population, and the sample characteristics differed from population parameters due to sampling error.

In sum, the manipulated parameters were:

- predictor distribution type: truncated-normal or gamma (tn vs. gamma)
- predictor proportion true-zero (.1, .5, .9)
- reliability of the predictor (via trial number: 40, 80, 160, 320 that defined the SD of the normally distributed error)
- reliability of the criterion (.1, .3, .5)
- slope (0.25, 0.5, 1)
- sample size (50, 150, 500)

Constants:

- The variance of the normal distribution before truncation was constant (sigma_x_nd=1).
- The regression intercept was fixed to zero.
- In each cell of the simulation, there were n_sim=1000 samples. (todo: n_sim anpassen)

7.2.1 Reliability of the Predictor

The reliability of the predictor was calculated by simulating SDT-model trials. Based on the reliability and the variance of the true predictor values, we calculated the variance of the error of the predictor and used it in the simulation below to simulate normally distributed error of the same size as the error resulting from the SDT model.

7.2.2 EiV Regression Estimation

In the EiV regression, obtaining p-values involves computations that occasionally fail, yielding missing values (specifically, the standard errors of the regression parameters could not be estimated). This problem was reduced but not eliminated by using different starting values for the EiV regression algorithm. If the algorithm did not succeed after trying out 10 different starting values, a new sample was drawn and EiV regression was again attempted; and this procedure was repeated until a sample was obtained for which EiV regression returned the standard error estimates necessary for computing p-values.

This replacement procedure did not affect the results. We ran a simulation with missing values, and compared the results to those of a simulation in which missing values were replaced. The results did not differ meaningfully: Across all dependent variables, all the predictors involving the replacement factor had very small effects (ges$<$.001).

7.3 Main Results

7.3.1 Intercept - Standard Regression

7.3.2 Intercept - EiV Correction

7.3.3 Alpha Inflation - Standard Regression

7.3.4 Alpha Inflation - EiV Correction

7.3.5 Slope - Standard Regression

7.3.6 Slope - EiV Correction

7.3.7 Slope Power - Standard Regression

7.3.8 Slope Power - EiV Correction

7.4 Additional Results

7.4.1 Alpha Inflation - EiV Correction: Alpha Level

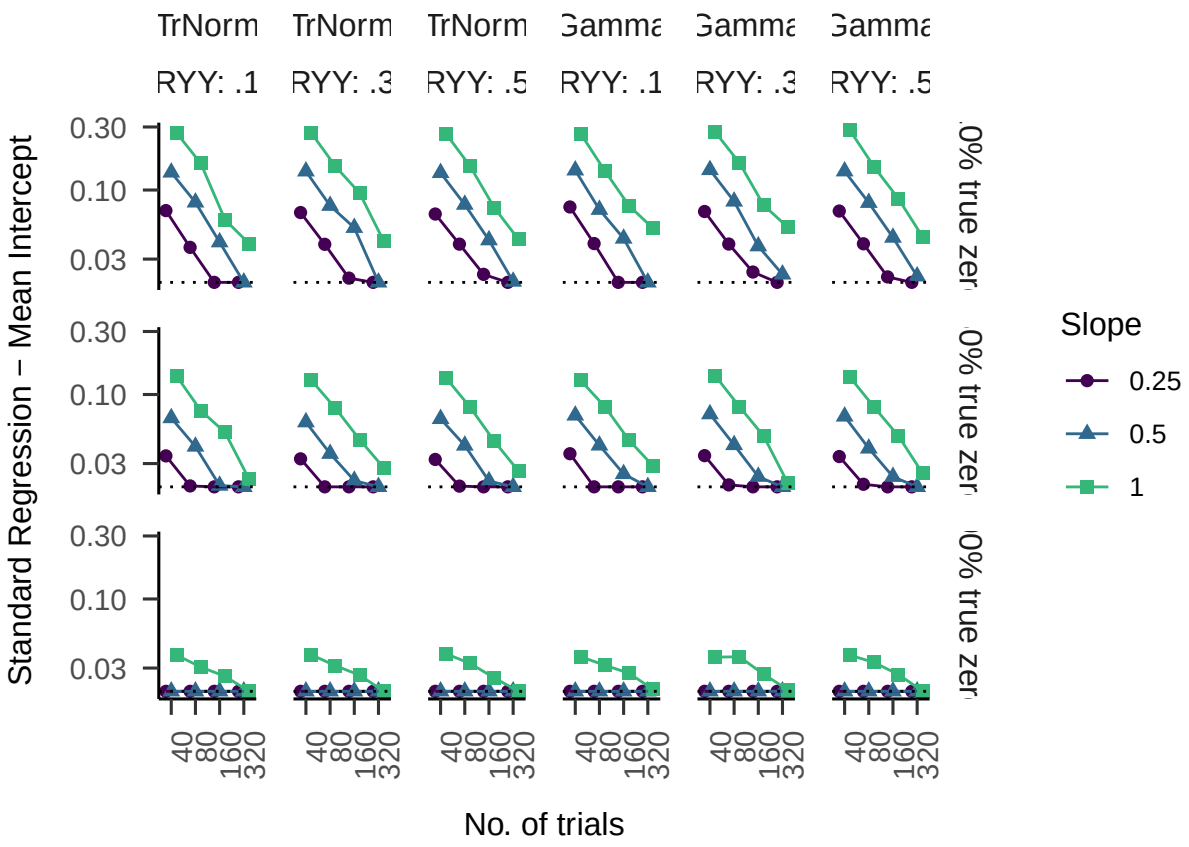

Figure 23: Intercept estimates (standard regression; for N = 150) as a function of measurement error (n_trials) and slope, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

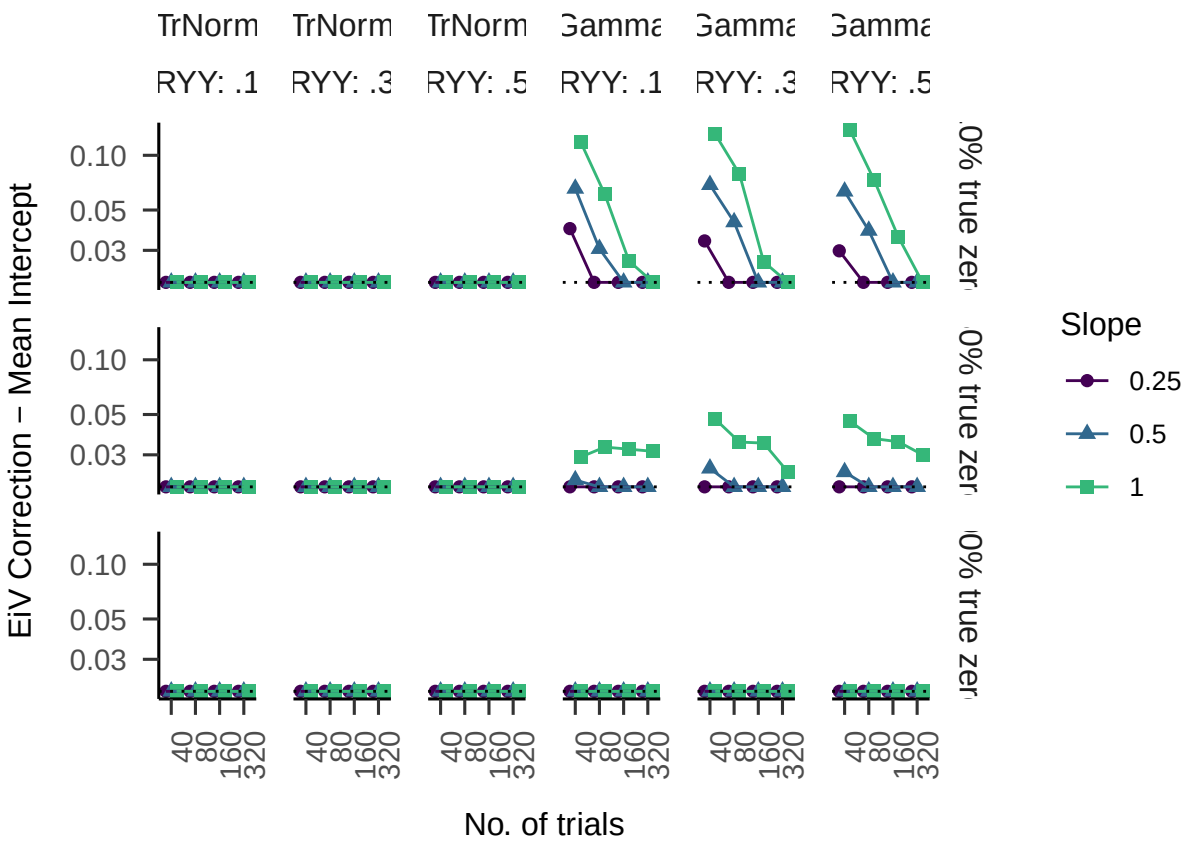

Figure 24: Intercept estimates (EiV regression; for N = 150) as a function of measurement error (n_trials) and slope, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

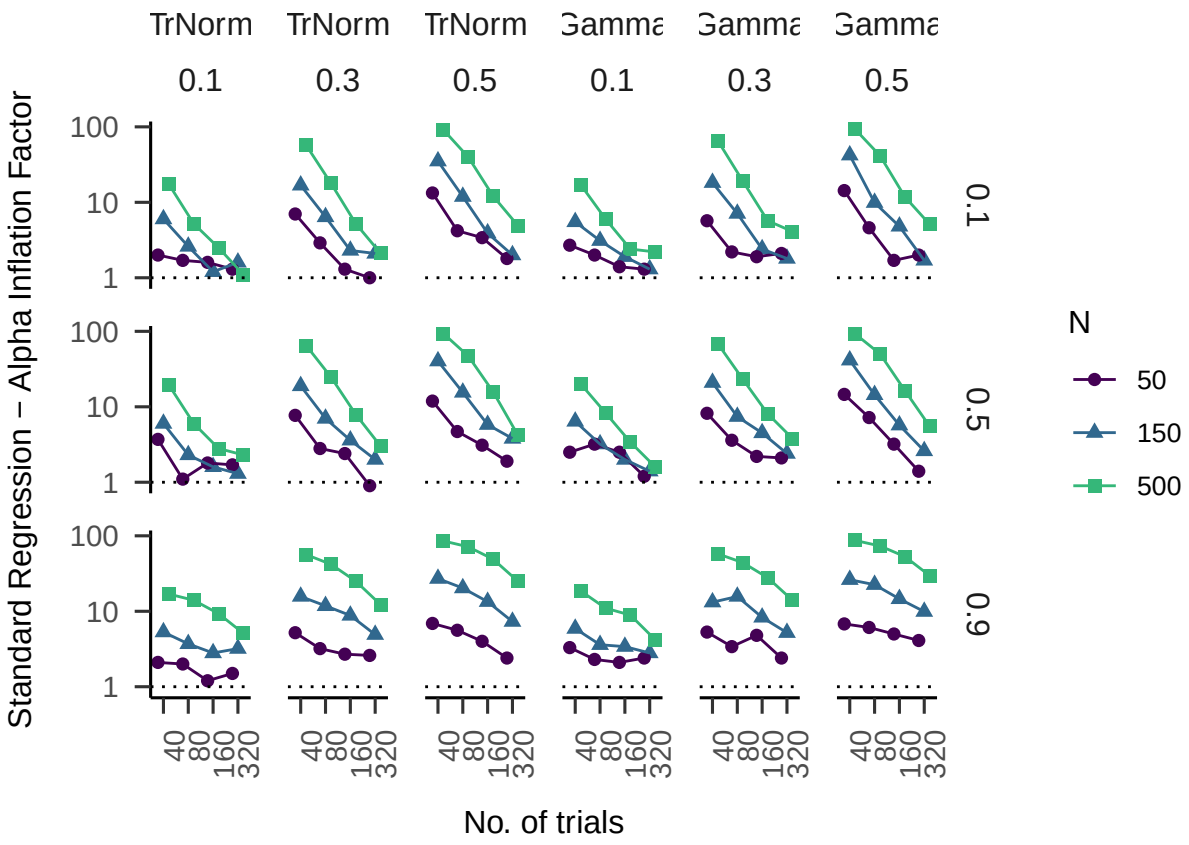

Figure 25: Inflation factors of the alpha error rate for the intercept test against zero (i.e., observed error rates divided by nominal alpha level; standard regression; for one-sided alpha = .01 and slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

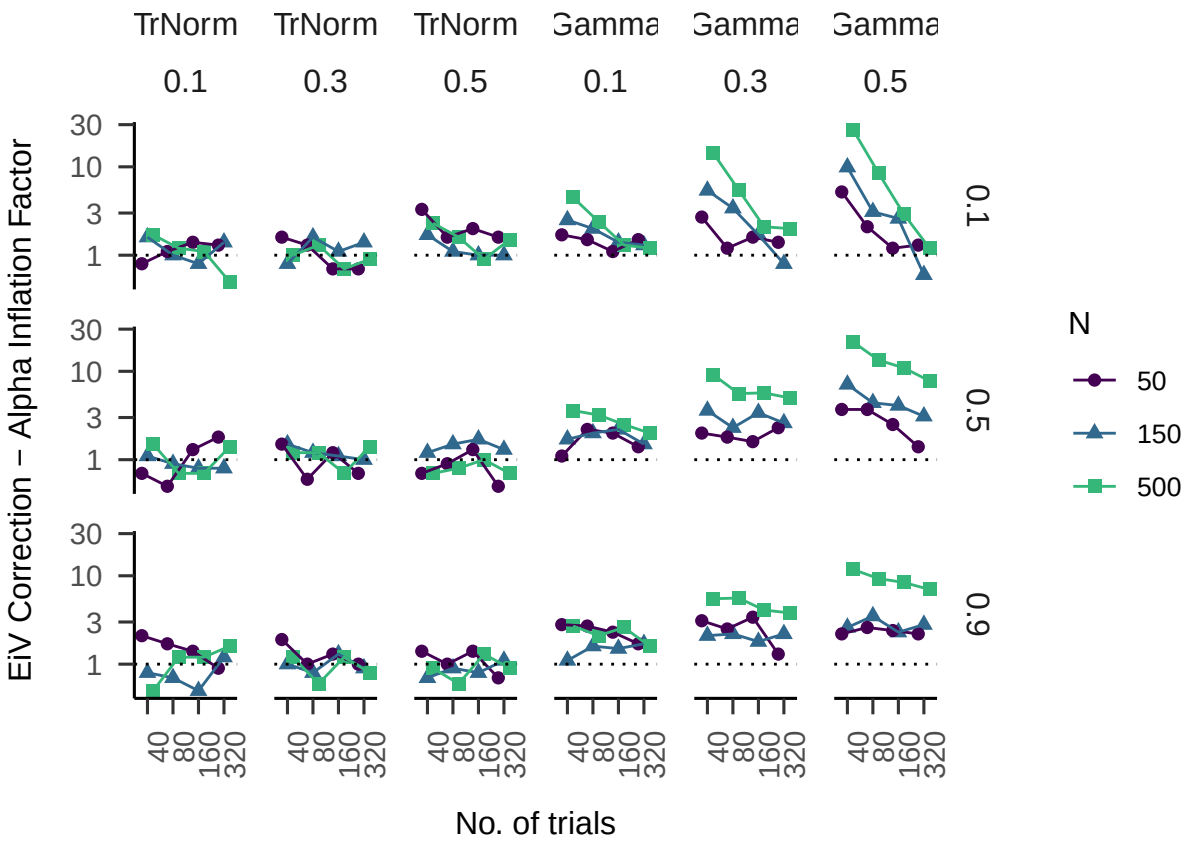

Figure 26: Inflation factors of the alpha error rate for the intercept test against zero (i.e., observed error rates divided by nominal alpha level; EiV regression; for one-sided alpha = .01 and slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and true-zero proportion (rows). Data values were plotted on a logarithmic (base 10) scale to accommodate the wide range of values.

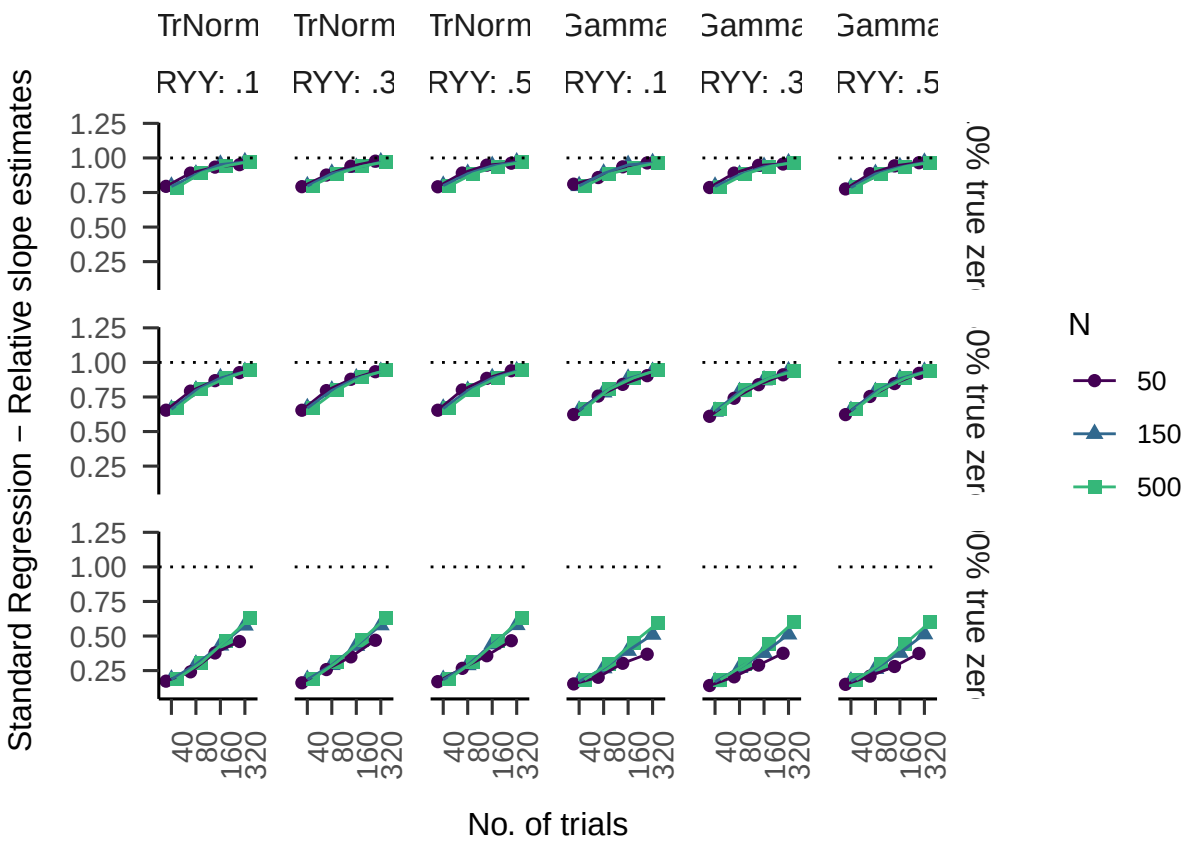

Figure 27: Relative slope estimates (i.e., mean observed slopes divided by true slope; standard regression; for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

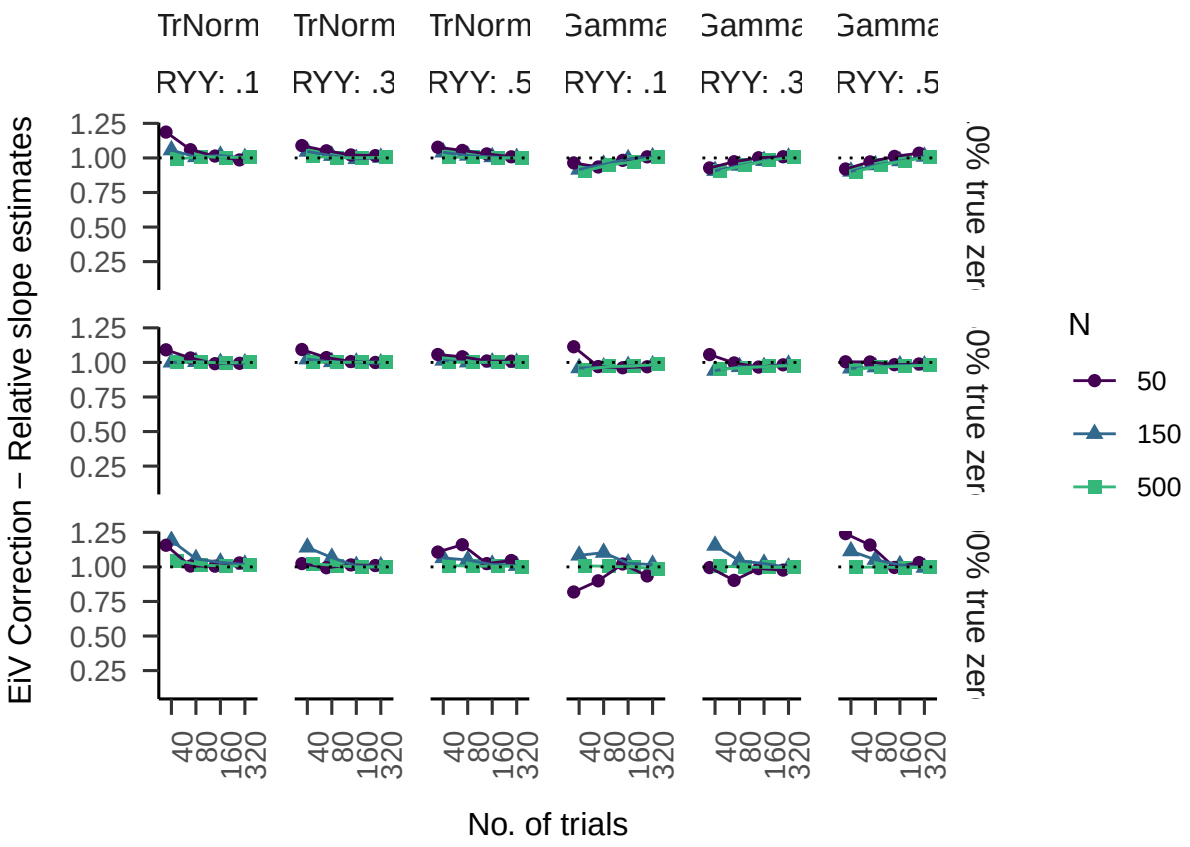

Figure 28: Relative slope estimates (i.e., mean observed slopes divided by true slope; EiV regression; for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

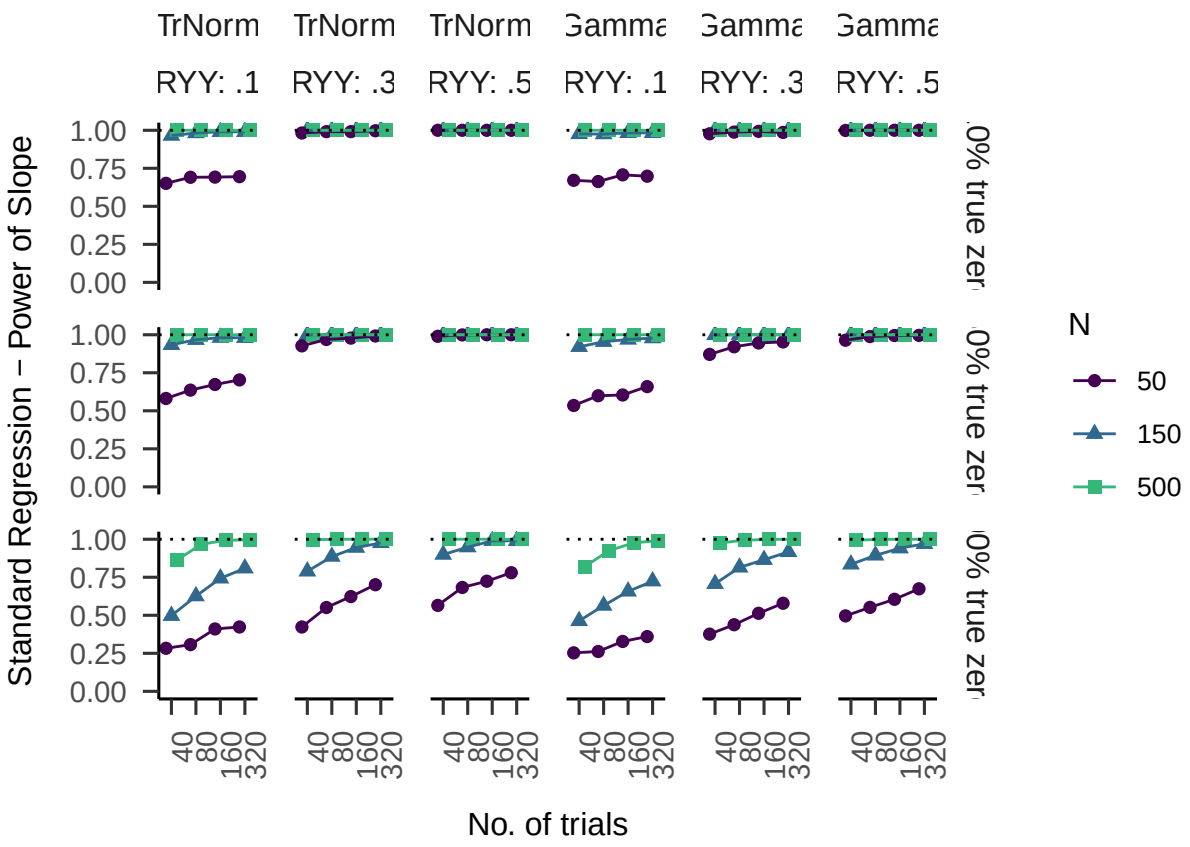

Figure 29: Power of the intercept test (i.e., proportion of results significant at one-sided alpha=.05; standard regression, for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

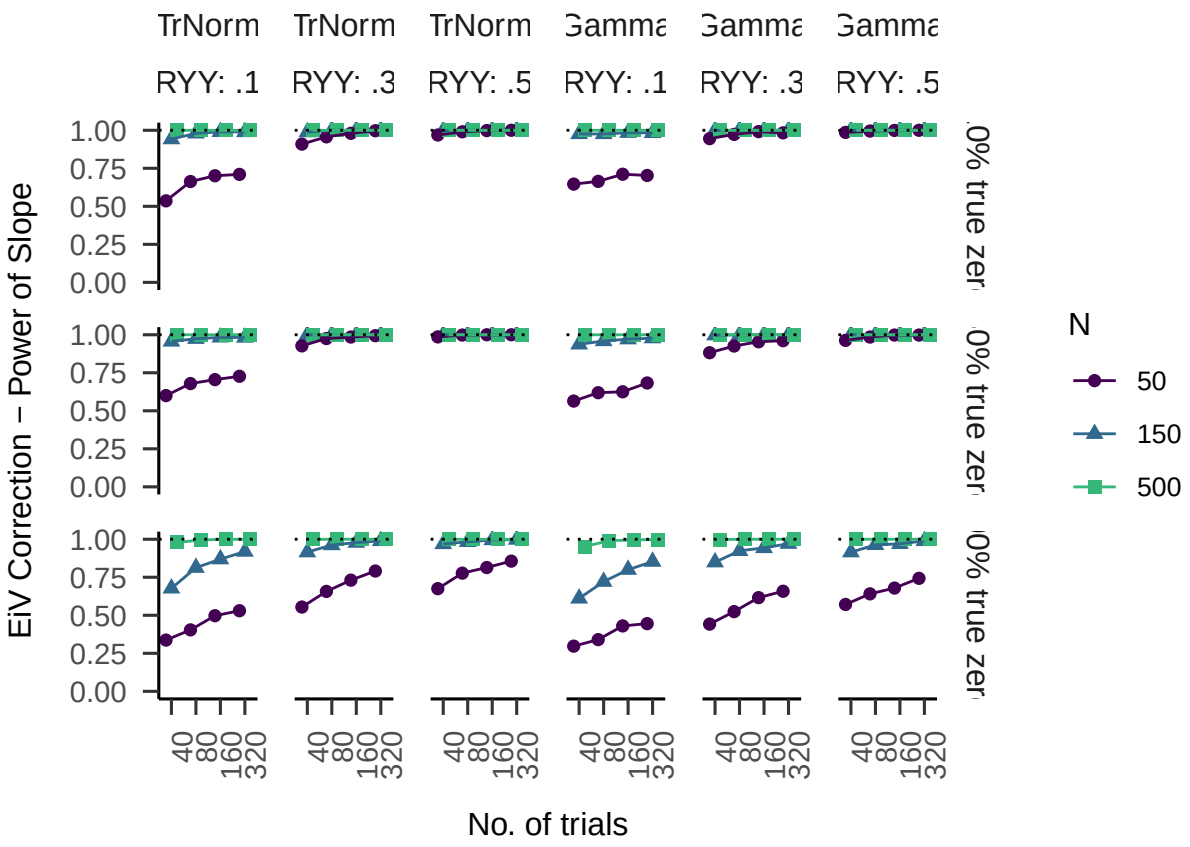

Figure 30: Power of the intercept test (i.e., proportion of results significant at one-sided alpha=.05; EiV regression; for slope = 1) as a function of measurement error (n_trials) and sample size, split by distribution and R_YY (columns), and proportion true-zero (rows).

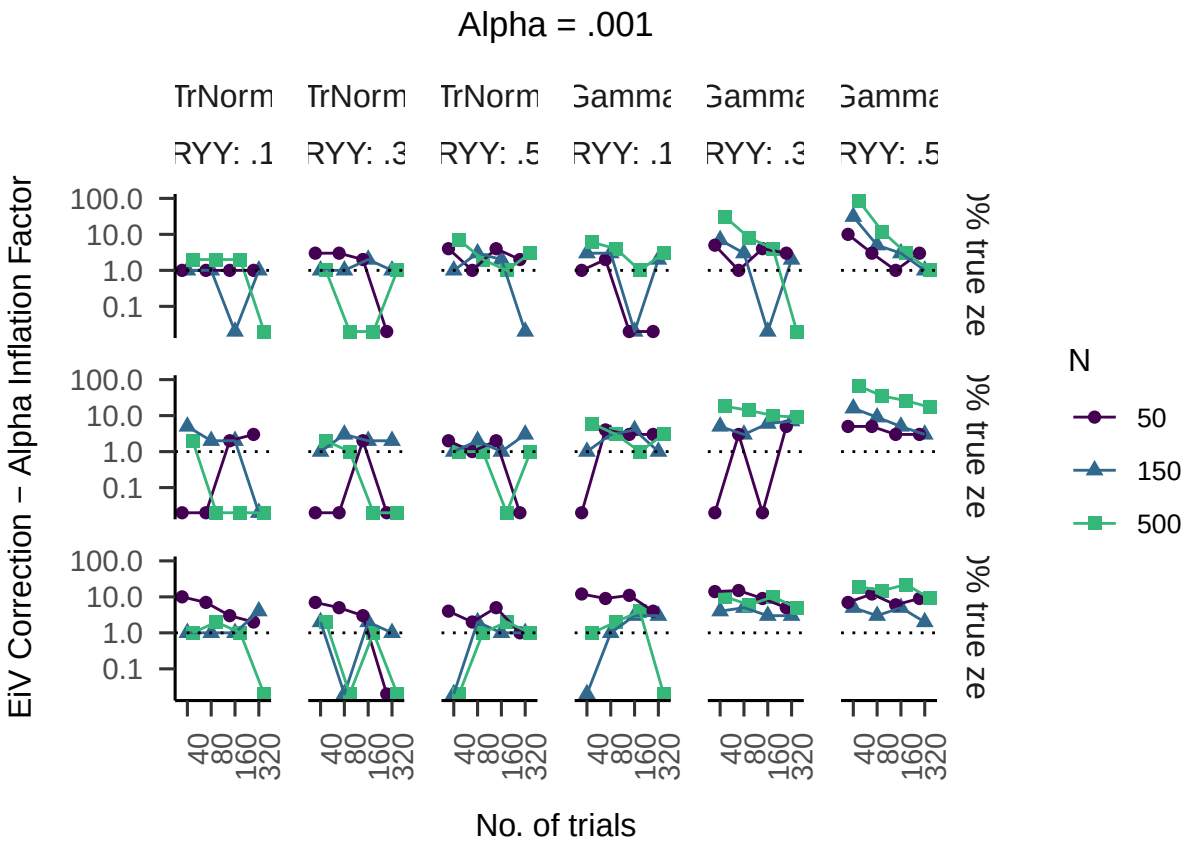

Figure 31: Inflation factor increased with decreasing alpha level.

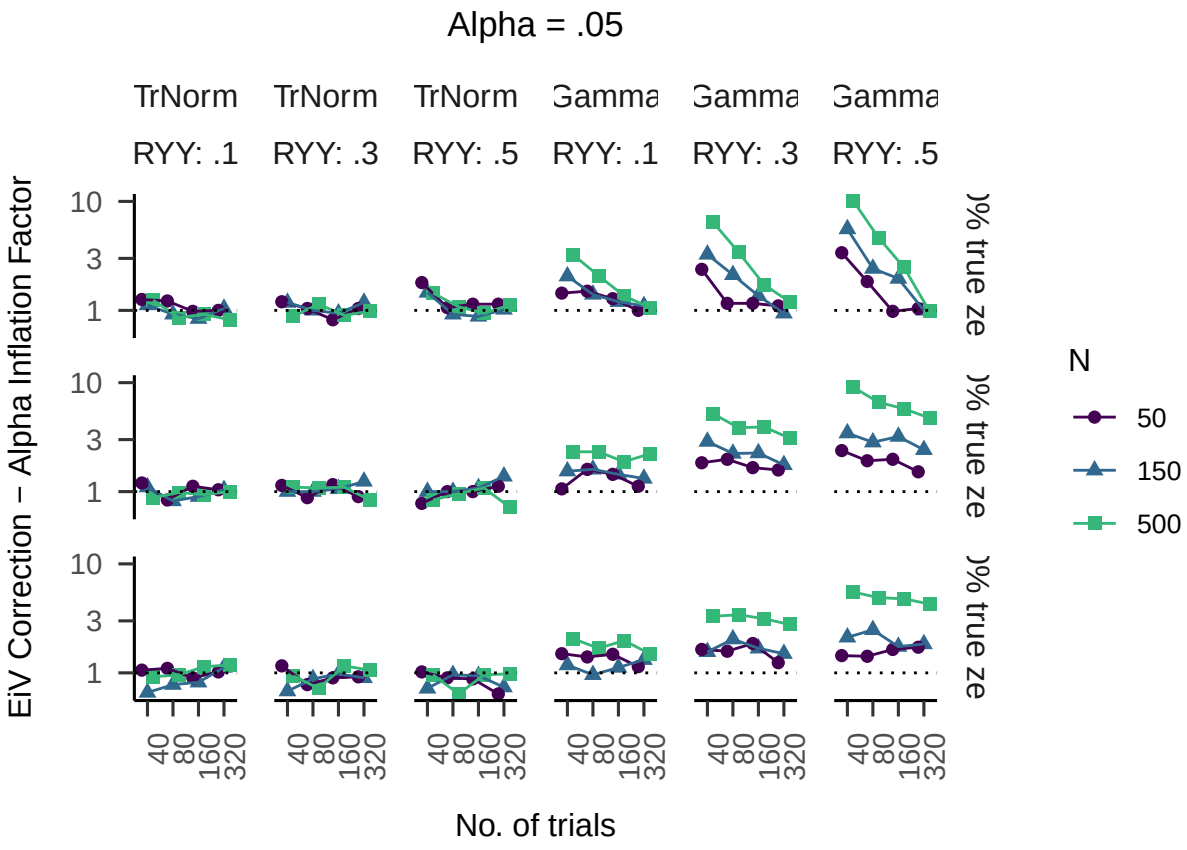

Figure 32: Inflation factor increased with decreasing alpha level.