ML2 vs. ML5 mystery

March 20, 2018

Questionnaire differences

We identified three minor differences in the design of the questionnaire that might have contributed to the discrepant results (Table 1, rows 1-3). First, the endpoints of the scale measuring perceived likelihood ranged from 1-10 in ML2 but ranged from 0-10 in ML5 and in the original study. Second, the questionnaire permitted subjects to skip questions in ML2, but not in ML5, potentially leading to systematic differences in characteristics of self-selected subjects; the original study was conducted on paper, so likely did not prevent subjects from skipping questions. Third, in ML2, the questionnaire was embedded in a roughly 30-minute series of experiments presented in a randomized order, which was not the case in ML5.

To investigate whether these differences in questionnaire design might have contributed to the discrepant results, we compared results between each study's sample of Amazon Mechanical Turk subjects in the United States. To the extent that these samples are directly comparable, any effects of the different questionnaire designs would likely produce differences in results between the MTurk samples. In contrast, the two samples estimated nearly identical point estimates (raw mean difference = 0.21, n = 340 in ML2 vs. 0.21, n = 2973 in ML5)[1]. Additionally, the extent of missing data was negligible in ML2 (0.5%), and their analyses of possible order effects suggested **XXX.**

Analysis differences

Additionally, the two studies used slightly different statistical analyses to aggregate data across sites: ML2 used an independent-samples t-test allowing for heteroskedasticity, while ML5 used a linear mixed model with model-based standard errors that assumed homoskedasticity. The original study used a one-way ANOVA (which assumes homoskedasticity). We therefore reanalyzed all subject-level data for primary analysis sites in both studies using identical statistical analyses using three approaches[2]:

1. To reproduce the analysis approach in ML5, we used a linear mixed model (**LMM**) to regress perceived likelihood on fixed effects of tempting fate, the interaction of tempting fate with study (ML5 vs. ML2), exchangeable random intercepts and slopes by site, and with normal errors. Inference used model-based standard errors, so assumed homoskedasticity.
2. To avoid parametric assumptions[3], we used generalized estimating equations (**GEE**) with a working independent correlation structure to regress perceived likelihood on tempting fate. Inference used robust standard errors, so made no assumptions about how subjects were correlated within sites or otherwise.
3. To reproduce the analysis approach in ML2, we computed raw mean differences (**RMD**) between tempting-fate conditions for each study separately, ignoring correlation of subjects within sites. We estimated standard errors for each as in the independent-samples Welch t-test conducted for ML2. We then tested the null hypothesis that the RMDs reflected the same population difference in both studies using the fact that, under the null:

$$\frac{\widehat{\Delta}_{ML2} - \widehat{\Delta}_{ML5}}{\sqrt{\widehat{SE}^2_{\widehat{\Delta}_{ML2}} + \widehat{SE}^2_{\widehat{\Delta}_{ML5}}}} \approx N(0,1)$$

[1] Of course, statistical inference on these point estimates differed due to the difference in sample sizes.

[2] Each of these models is saturated, so is unbiased for the point estimate; thus, differences would emerge primarily in statistical inference.

[3] Although GEE is in general semiparametric rather than nonparametric, here the model is saturated due to the categorical predictors, so is effectively nonparametric.

where $\widehat{\Delta}_{ML2}$ and $\widehat{\Delta}_{ML5}$ denote the estimated RMDs in ML2 and ML5 respectively.

All three approaches yielded strong evidence for persistent discrepancies between the studies (Table XXX).

Table 1: Difference between ML5 and ML2 in average main effect estimates (as raw mean differences) in primary analysis sites under varying statistical assumptions

Method	Allows correlation within sites	Assumes homoskedasticity	Estimate	95% CI	p-value
LMM	Yes, with normal site effects	Yes	-0.45	[-0.82, -0.08]	0.02
GEE	Yes, without assumptions on structure	No	-0.45	[-0.83, -0.07]	0.02
RMD	No	Yes	-0.45	[-0.78, -0.12]	0.01

Sampling frame differences

The most conspicuous protocol difference involved the sampling frame used in primary analyses: ML2 analyzed undergraduates at a variety of colleges and universities in the United States and abroad, while ML2 analyzed undergraduates at similar sites. We speculated that results might align more closely if we constructed more directly comparable sampling frames, so we redid all three analyses above (LMM, GEE, and RMD), but expanding the sampling frame to all sites that collected data for each study. Thus, the expanded sampling frames for both ML2 and ML5 included subjects collected in one or more online samples, undergraduates at a small number of universities that would be classified as "similar" by ML5, and undergraduates at a larger number of domestic or foreign universities that would be classified as "dissimilar" by ML5.

Table 2: Difference between ML5 and ML2 in average main effect estimates (as raw mean differences) in all sites under varying statistical assumptions

Method	Allows correlation within sites	Assumes homoskedasticity	Estimate	95% CI	p-value
LMM	Yes, with normal site effects	Yes	-0.27	[-0.52, -0.02]	0.03
GEE	Yes, without assumptions on structure	No	-0.29	[-0.46, -0.13]	0.00
RMD	No	Yes	-0.29	[-0.47, -0.12]	0.00

Lastly, we investigated whether outlying sites might have strongly influenced results in one or both studies. These outliers could arise, for example, from idiosyncrasies of protocol administration, subject characteristics, or errors in data collection or analysis. From a visual inspection of the forest plots (Figure 1), none of the primary analysis sites in either study appeared to be an outlier. Among all sites, Eotvos Lorand University (in ML5) may have been a modest outlier, but because this site estimated a large positive effect, its presence would, if anything, have reduced rather than exacerbated the discrepancy between studies. Outlying sites therefore do not appear to account for the discrepancy.

A combined estimate of tempting-fate effects

We combined all data from both studies to arrive at an updated estimate of the average effect of tempting fate across the diverse samples included in both studies. We again fit LMM and GEE models to primary analysis sites from both studies, omitting the interaction of tempting fate with study to estimate an average effect. The LMM model estimated a tempting-fate effect of 0.48 on the raw mean difference scale (95% CI: [0.34, 0.61]; $p = 8 \times 10^{-12}$). As expected, the GEE yielded a similar estimate (0.46) with more conservative inference (95% CI: [0.32, 0.61]; $p = 4 \times 10^{-10}$). The naïve Cohen's d effect size was an estimated 0.19 (95%

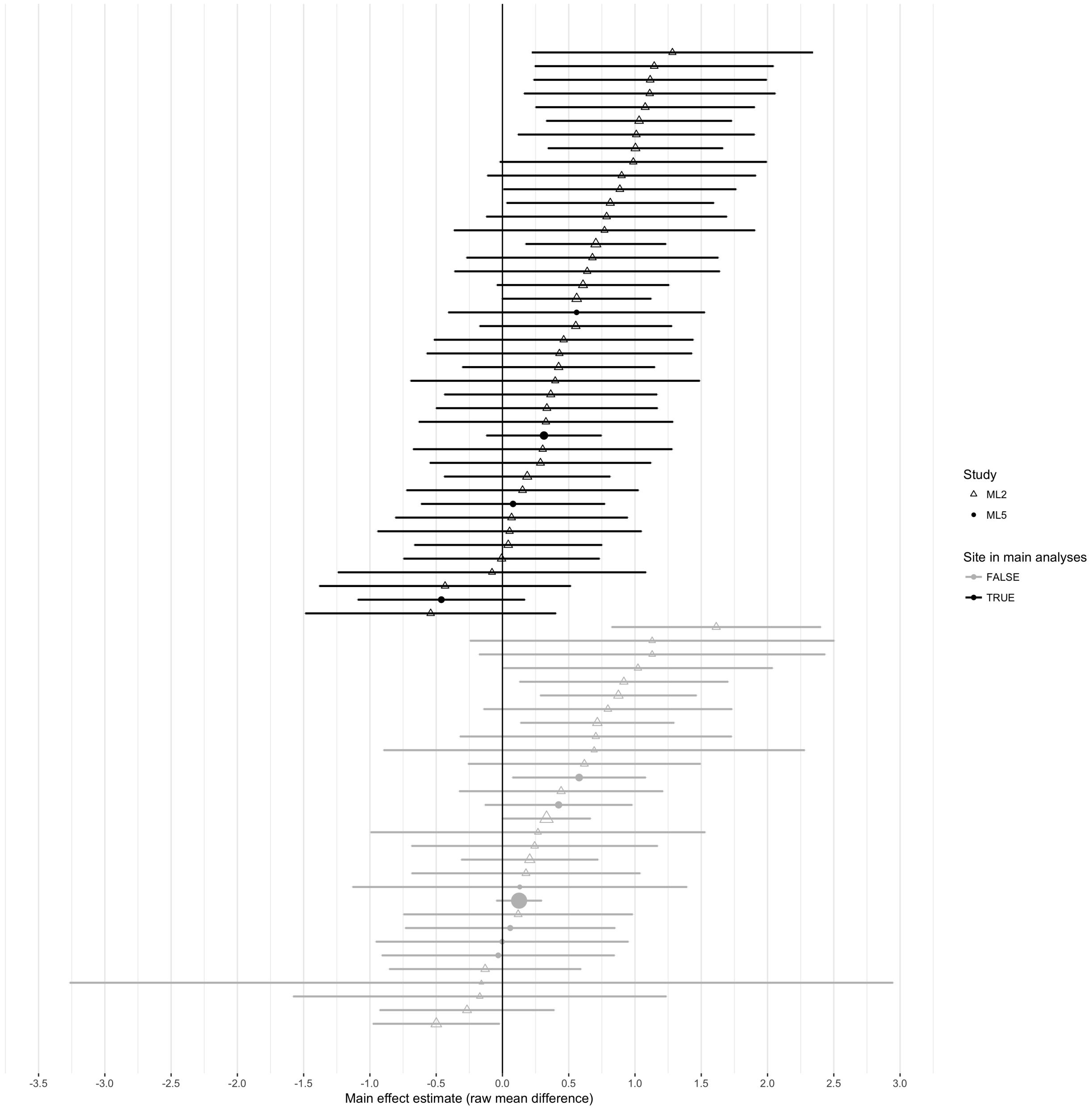

Figure 1: Main effect estimates for sites used in primary analyses (black) and those not used in main analysis (gray) for ML2 (triangles) and ML5 (circles). Plot symbol size is inversely proportional the estimated within-study variance.

CI 0.25, 0.14). To assess consistency between this pooled estimate and the estimate of the original study, we estimated that, if the original study were consistent with all replications, then the probability of observing a point estimate in the original study at least as extreme as that actually observed would be approximately $P_{orig} = 0.28$ (CITE). This method assumes normally distributed true effects, which appeared reasonable here, and accounts for the observed heterogeneity across sites.

Alternatively, including all sites that collected data for either study had little effect on point estimates or inference. The LMM model estimated a tempting-fate effect of 0.41 on the raw mean difference scale (95% CI: [0.3, 0.52]; $p = 4 \times 10^{-13}$). The GEE yielded a similar estimate (0.33) but with more conservative inference (95% CI: [0.19, 0.46]; $p = 3 \times 10^{-6}$). The naïve Cohen's d effect size was an estimated 0.19 (95% CI 0.25, 0.14). As a metric of consistency with the original study, we estimated $P_{orig} = 0.25$; however, this must be interpreted cautiously in light of possible departure from normality.

FIX DIRECTION OF COHEN'S

Conclusion

- Discuss combined point estimates and their inference
- These estimates should be interpreted cautiously because blahblahblabh.
- Point estimate much smaller than original but still consistent because of original's small sample size
- Would like to see future literature use other scenarios…