Pre-analysis Plan for Measuring "Ethnic Effects" via Social Psychology Paradigms and Behavioral Games

Chad Hazlett Daniel N. Posner Ashley Blum*

February 1, 2018

Overview

The objective of the project described in this pre-analysis plan is to compare the size of the "ethnic effect" measured via standard social psychology paradigms and behavioral economics games. By "ethnic effect," we refer to the difference in behavior under one condition (involving behavior towards a coethnic) and another condition (involving behavior towards a non-coethnic).[1] Generally it is these measures of ethnic effects of preferences that we imagine others possibly using in their own research, should they prove to be useful here. The social psychology paradigms we employ are the Affect Misattribution Procedure, the Face Affect Attribution Task, and the Weapon Misidentification Task. The behavioral economics games we employ are the Dictator, Public Goods, and Choose-Your-Dictator Games. Bridging notation across disciplines, we use the term "tasks" to refer to each of these experimental procedures, encompassing both the psychology "paradigms" or the behavioral economics "games".

In addition, we employ a fear-prime randomized at the subject level. This is designed as a validity probe, based on the premise that if these measures pick up fear of the other group, then priming fear may amplify the effect sizes.[2]

Our sample is drawn from Nairobi, Kenya and includes members of the Kikuyu and Luo ethnic communities. Participants in the tasks are primed to think about and/or paired with others whose ethnic identity is primed by referring to either a hometown or a province that is widely associated with members of these two groups. The match betweeen the participant's own ethnic group membership and that of the person they are primed to think about/are paired with determines whether the particular round of the task is coded as "coethnic" or "non-coethnic."

*Departments of Statistics and Political Science, University of California, Los Angeles.

[1] The exception is behavior in the Choose-Your-Dictator game (described below), which comes instead in the form of a revealed preference for one group vis-a-vis the other (and is meaningful only across participants).

[2] We acknowledge several limitations of this approach from the outset, described below. A consequence is that even if our measures pickup "fear," the fear prime may not modulate them.

Altogether there are five main branches of the analysis:

Branch 1 Primary analyses of the “ethnic effect” estimates produced by each task (or the ethnic preference measure in the Choose-Your-Dictator game), the relationships across tasks on these measures, and the distribution of effect estimates across individuals.

Branch 2 Analysis of whether, for each task, these ethnic effects are modulated by the fear prime.

Branch 3 Further analysis of ethnicity effects, involving various sub-group effect estimates or correlations with various observables, heterogeneity, etc.

Branch 4 Follow-up study using outcomes from phone calls after the experimental procedure, measuring attitudes toward the election which we expect to be ethnically informed.

Branch 5 Additional analyses that, whether anticipated or not at present, are exploratory. These additional analyses will be performed in a subset of the data followed by confirmation/testing in a held-out sample.

Below we describe each of these branches of the analysis in turn.

Branch 1: Ethnicity effect estimates, detectability, and sensitivity

In this branch of the analysis, we ask “Can we detect any effect of ethnicity using these measures?” and “How strongly do each of these tools pick up any such effect?”

Social Psychology Tasks

The psychological tasks are:

- AMP^{town}: the Affect Misattribution Procedure, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.
- AMP^{prov}: the Affect Misattribution Procedure, where the primes are the names of the two provinces most clearly assoicated with the Kikuyu and Luo.
- FA^{slow}: the Face Affect Attribution task, presented at a slow rate, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.
- FA^{fast}: the Face Affect Attribution task, presented at a faster rate, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.
- WMT: the Weapon Misidentification task, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.

The AMP, FA, and WMT all have a similar structure: each contains numerous trials, each of which consists of a cue or prime, followed by a target image, followed by a response. Moreover, the quantities of interest are similar for each: how much more likely are you to give a particular response when primed to coethnicity than when primed to non-coethnicity.

The quantities of interest are thus priming effect estimates, taking the form of differences-in means. These will first be constructed at the individual level (averaging across trials) to obtain individual level measures. Specifically, we define the following quantities for each individual i. For the AMP-hometown and the AMP-province tasks, the quantities of interest are:

$$
\begin{aligned}
AMP_i^{town} &= Pr[\text{pleasant}|\text{coethnic}] - Pr[\text{pleasant}|\text{non-coethnic}] \\
AMP_i^{prov} &= Pr[\text{pleasant}|\text{coethnic}] - Pr[\text{pleasant}|\text{non-coethnic}]
\end{aligned}
$$

where *pleasant* is the event that the participant pressed the button indicating that they thought the image they were shown was "pleasant" (rather than "unpleasant"), and ($coethnic$ or $non - coethnic$) refer to the cued ethnicity (through the particular town or province) and it's match to the participant.

For the Face Affect Attribution (FA) task, we have a slow version and a fast version, with quantities of interest:

$$
\begin{aligned}
FA_i^{slow} &= Pr[\text{happy}|\text{coethnic}] - Pr[\text{happy}|\text{non-coethnic}] \\
FA_i^{fast} &= Pr[\text{happy}|\text{coethnic}] - Pr[\text{happy}|\text{non-coethnic}]
\end{aligned}
$$

where *happy* refers to pressing a button indicating the target was thought to be happy (rather than angry), and the cues ($coethnic$ or $non - coethnic$) refer to the cued ethnicity and its match to the participant.

Finally for the Weapon Misidentification Task (WMT), we have:

$$WMT_i = Pr[\text{non-weapon}|\text{coethnic}] - Pr[\text{non-weapon}|\text{non-coethnic}]$$

where $non - weapon$ refers to pressing a button indicating that the target was thought not to be carrying a weapon, and the cues again may be $coethnic$ or $non - coethnic.$

We note that each of these has the form,

$$\text{EthnicEffect}_i = Pr[r_i|c_i] - Pr[r_i|\bar{c}_i]$$

where r is one of the (two) possible response types for a given task (generally the "positive" one, by convention), c is the cue to coethnicity, and $\bar{c}$ is the cue to non-coethnicity. Each of these quantities will be estimated for person i using simple difference-in-means. That is, the probabilities will be replaced by the corresponding conditional averages, as in

$$\text{EthnicEffect}_i = \frac{1}{N_c}\sum_i r\mathbb{1}_c - \frac{1}{N_{\bar{c}}}\sum_{i=1} r\mathbb{1}_{\bar{c}}$$

where r equals 1 for the "positive" response type and 0 for "negative" response type. Finally, when averaging these quantities across the individuals in the study, to indicate the group average we write $\overline{\text{EthnicEffect}} = \frac{1}{N}\sum_{i=1}^{N} \text{EthnicEffect}_i$.

Recall that the outcome in each trial of the Face Affect Misattribution Task is whether participants believe the person they are viewing is angry; the outcome in each trial of the weapon misidentification task is whether they think the person they are seeing is holding a weapon. Any effect of ethnicity on these outcomes thus suggest *prima facie* evidence that ethnicity is connected to "fear" or "threat" perception, as it is difficult to explain the observed relationship otherwise. However, further questions of construct validity are examined below.

In addition to the analyses involving the AMP described above, we also have participants undertake an additional task, the AMP^{univ}, where the primes are words of near universal positive or negative appraisal (e.g. sunshine, disease). This is the version of the AMP that is commonly used in U.S.-based laboratories. Our purpose in using it in the context we study is to gauge whether the AMP operates similiary in our (quite different) setting, which is characterized by low literacy and low familiarity with computers and keyboards.

The analyses to be conducted with these measures are described below after introduction the analogous behavioral game measures.

Behavioral Economics Tasks

The behavioral economics tasks can be understand in terms of the same structure and objects of inference. The tasks are:

- DG: the Dictator Game, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.
- PG: the Public Goods Game, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.
- CYD: the Choose-Your-Dictator Game, where the primes are the names of hometowns widely known to be predominantly Kikuyu or Luo.

For the Dictator Game (DG), the question is how the "contribution" of money made by the participant to the other player depends upon the other player's ethnicity, which is inferred from the player's hometown. Our measure is thus

$$DG_i = \mathbb{E}[\text{contribution}_i|\text{coethnic}] - \mathbb{E}[\text{contribution}_i|\text{non-coethnic}]$$

where the *contribution* is simply the amount of money given to the other player, the *coethnic* or *non-coethnic* conditions are determined by the hometown name of the other player combined with the ethnicity of the participant.

This difference does not tell us whether, compared to some baseline, people show "in-group favoritism" or "out-group hostility." To inform this, we can compare each of the quantities above – $\mathbb{E}[\text{contribution}_i|\text{coethnic}]$ and ($\mathbb{E}[\text{contribution}_i|\text{non-coethnic}]$ – to the mean contribution in a condition in which participants do not know the ethnicity of their counterpart (what we call the "non-profiled" condition). However, we save that analysis as exploratory.

For the Public Goods Game (PG), participants play with three-person groups that are either homogeneously coethnic, homogeneously non-coethnic, or a mix. For present purposes, we are less interested in the mixed condition, for which we have no strong theoretical expectation. Our measure is:

$$PG_i = \mathbb{E}[\text{contribution}_i|\text{coethnic}] - \mathbb{E}[\text{contribution}_i|\text{non-coethnic}]$$

where *contribution* is the contribution made to the group, the *coethnic* condition indicates that the two other players are of the same ethnicity as the participant, and the *non-coethnic* condition indicates that the two other players are of the other ethnicity (both inferred by the hometowns of the other players combined with the ethnicity of the participant in question). We further note that the "mixed" condition will be analyzed as part of our exploratory analysis.

Again, a non-profiled version of the game will be used in exploratory analyses as a "baseline" to help determing is any ethnic effect we see is "in-group favoritism" or "out-group hostility" relative to the case where ethnicity is not known.

Results from the Choose-Your-Dictator Game (CYD) are analyzed somewhat differently. While the DG and PG produce the same type of ethnic effect measure as the prior tasks, the CYD produces instead two coarse measure of a realized preference for each participant. The first, "non-profiled," trial of the CYD asks participants whether, in a context in which the Dictator does ***not*** have knowledge of the participant's own ethnic group membership, they would prefer to have a coethnic or non-coethnic play the role of Dictator in a DG in which they are the receiver. They are also given the option of indicating that they are indifferent. The outcome takes a value of 1 if they prefer to play with the coethnic, and 0 if they prefer to play with the non-coethnic or are indifferent.

The second, "profiled," trial of the CYD repeats the question, but now in a context where the participant is told that the Dictator will be aware of the participant's ethnicity. The participant can, again, choose to play with a coethnic, a non-coethnic, or say they are indifferent. The outcome again takes a value of 1 if they prefer to play with the coethnic, and 0 if they prefer to play with the non-coethnic or are indifferent. Both the "non-profiled" and "profiled" versions of the CYD are one-off measures rather than differences between conditions within individuals, so they relay somewhat less information on the individual level. However, they may nevertheless prove correlated with other measures and characteristics across individuals. As with the other measures, we describe statistical tests involving the behavioral measures below. The two resulting measures are $CYDp_i$ and $CYDnp_i$ and are simply the responses on these trials – 1 if i prefers to play with the coethnic, 0 if i prefers to play with the non-coethnic or is indifferent.

Rejection Criteria

Before describing our analysis and inferential procedures, we commit to a set of data rejection criteria we describe below. As the challenges of introducing computerizing experiments in the population we study are especially acute, such rejection criteria are important for data quality control and to avoid finding effects that are widely biased by a share of the sample not actually understanding or able to complete the experiments. These rejection criteria were chosen after we had looked at the distribution of several variables (literacy test, Raven's matrix results, and reaction times in the psych tasks), but only on a randomly selected half of the data (the half that will be used as the "exploratory set" in split-sample exploratory analyses described in Branch 5, below). Beyond these three variables, we have blinded ourselves to any other outcomes and conducted no analyses.

The first rejection criterion is ***reaction time*** (RT). We will:

- Cut any *trials* on which the $RT < 200$ms, as this is implausible and will typically indicate inappropriate button pressing.
- We also register our intention to conduct a robustness test in which we drop *participants* with more than some portion of trials below 200 ms, perhaps 20%.

The second criterion is ***literacy***. We propose to:

- Keep all participants in the main analysis regardless of literacy
- We also register our intention to conduct a robustness test in which we re-run analysis but dropping at participants with only 0 to 2 correct answers in the six-question literacy test.

The third criterion is the score on ***Raven's matrices***. We will:

- Keep all participants in the main analysis regardless of their score in the Raven's matrix test
- We also register our intention to conduct a robustness test in which we drop those participants who get two or fewer correct. We are particularly interested in knowing if (a) the behavioral economics games in particular work better among those with higher scores, and if (b) the correlation between the psychological and economics tasks is higher when participants' performance on the Raven's test is better.
- We register our intention to conduct further exploratory analyses beyond dropping those with two or fewer correct, depending on the results of the above questions.

We also register our intention to examine which participants are most likely to be following a pattern in their responses that may be indicative of failing to follow the task instructions. This involves regressing their responses on indicators for prior response patterns to determine if the pattern of responses is more structured (e.g. alternating, or repeating) than expected by chance. We will use this only in a post-hoc analysis to see if our results tend to be a function of the probability that people may be following a set pattern of responses rather then responding genuinely.

More broadly, since we are implementing many of our tasks in a novel context, we anticipate conducting a number of exploratory analyses (across different sub-groups or including different covariates) that go beyond the tests we explicitly register here. Those exploratory analyses (some of which are discussed in Branch 3, below) will be clearly labeled as such in our write-up.

Individual Means

Most analysis will be done using subject-level data. The matrices M_{psych} and M_{econ} will each contain one row per participant, and columns containing each participant's mean *EthnicEffect* measure for each task as described above. We will then construct a z score for each column/task against a comparison value of 0 – i.e. if X is one column of M, compute $z = \overline{X}/\frac{SD(X)}{\sqrt{(N)}}$. This will simply facilitate scale-free comparison.

The first question is whether these tests pick up any effect of ethnicity on average that is distinguishsable from zero. We will thus test hypotheses

$$\overline{AMP}^{univ} > 0 \tag{1}$$
$$\overline{AMP}^{town} > 0 \tag{2}$$
$$\overline{AMP}^{prov} > 0 \tag{3}$$
$$\overline{FA}^{slow} > 0 \tag{4}$$
$$\overline{FA}^{fast} > 0 \tag{5}$$
$$\overline{WMT} > 0 \tag{6}$$
$$\overline{DG} > 0 \tag{7}$$
$$\overline{PG} > 0 \tag{8}$$
$$\overline{CYD} > 0 \tag{9}$$

These tests will be done using two-tailed t-tests against null hypotheses of zero effect. Having constructed z for each task, we need only check if its absolute value is greater than 1.96 (or the corresponding critical value for the t distribution). While we have strongly directional hypotheses, we will use two-tailed test statistics regardless of our skepticism.[3]

We also register our intent to remark on the z scores for each task, as an indicator of sensitivity. The z score for each task is an estimate of how surely we can reject the null hypothesis of "no ethnic effect" for each measure. If one task gets us an ethnic effects that averages 5 standard errors (i.e. $z = 5$), we are picking something up with it much more surely than a test that has $z = 0.5$. This is effectively a signal-to-noise ratio for each task.To this end, we will obtain confidence intervals on the z-scores from each test by bootstrap, and produce a plot with the z-score and its confidence interval for the ethnic effect measured by each task.

[3] We note that these data have already been collapsed to the individual level – each individual's score is the unweighted average across their trials. With "one number per person" on each task, it is no longer a repeated measures scenario, and there is no need for cluster-robust standard errors or multi-level models.

Individual Level Effects and Distributions of Effects

Typically, a measure is only available once or several times per participant and is taken as a single draw of a random variable. While this is true of our behavioral games measures (DG, PG, CYD), our psych tasks each contain dozens of trials. This makes it possible to obtain useful information about individual level estimates for these tasks.

Individual Detection

In other work on the AMP (Hazlett & Berinsky, 2017), the priming effect is dramatic enough that even within single individuals there is often an effect statistically distinguishable from zero. We will similarly test for individual-level significantly non-zero estimates on all psych tasks:

$$\overline{AMP}_i^{univ} > 0 \quad \forall i \tag{10}$$

$$\overline{AMP}_i^{town} > 0 \quad \forall i \tag{11}$$

$$\overline{AMP}_i^{prov} > 0 \quad \forall i \tag{12}$$

$$\overline{FA}_i^{slow} > 0 \quad \forall i \tag{13}$$

$$\overline{FA}_i^{fast} > 0 \quad \forall i \tag{14}$$

$$\overline{WMT}_i > 0 \quad \forall i \tag{15}$$

Distributional Results

In addition, though not a statistical test, we register our intent to show the distribution of the individual-level estimates for each measure. Prior research with the AMP has shown that this can reveal groups of participants who vary widely on their response. Specifically, for each of the psychology tasks, we will show estimated distributions of the individuals level ethnic effect point estimates.

Individual Level Cross-Task Comparisons

A key aspect of our analysis is comparison across tasks. We plan to:

- Check the full correlation matrix, $cor([M_{psych}, M_{econ}])$. Remark on the correlations within psych task (upper left), within the econ tasks (lower right), and between the two groups (either off-diagonal block).

- Check scree plots for M_{econ} and M_{pscyh} and particularly the variance explained by the first dimension of each.

- Compare the first dimension of M_{econ} to that of M_{psych} by correlation.

- Check how much of M_{econ} can be explained by M_{psych} in terms of the multivariate R^2.

If elements of this analysis are re-run on subsets of the sample, this will be done under the exploratory analysis section of the results.

Branch 2: Fear prime modulation of ethnicity effects

In the second branch of the analysis, we test for a priming effect of fear modulation on the ethnic effects estimated in branch one. We note that this is just one line of argumentation we have for the validity of these tests as measures of fear or threat perception.[4] We also register that we anticipate several reasons why the fear-prime may not influence the ethnic effects even if the latter do measure a fear-related construct. For example, if the fear prime makes people more likely to give the fearful response under *both* coethnic and non-coethnic conditions, then we will not see a change in the ethnic effect. Alternativeley, a fear prime may not be sufficient to alter the deep and automatic processes on which these measures depend.

The first analysis we plan is to simply compare the ethnic effect sizes in the fear-primed versus the non-fear-primed group. For the 9 tasks (6 psych; 3 econ), this means 9 comparisons. For each we can (a) do a simple two-sample t-test on the ethnic effect estimates; and (b) plot separately the two relevant means that go into the effect estimate for each task (mean behavior on coethnic trials, mean behavior on non-coethnic trials), under both fear-primed and non-fear prime conditions, together with confidence intervals.

Second, we will also test if the entire distribution of individual ethnic effects, rather than just the mean, shifts under priming. For each task we will plot the distribution of ethnic effects under the fear prime and under the non-fear prime. We will use a Kolmogorov Smirnov test to determine if these distributions significantly differ across primes within each task.

Branch 3: Additional planned comparisons

Having checked whether the fear primes move the ethnic effect measures, we are also interested in how ethnic effects relate to various other covariates. Each such comparison is interesting either substantively (e.g. how does urban exposure relate to ethnic bias?), diagnostically (e.g. who should have an effect and who should not? does task ordering matter?), or in terms of concurrent validity (e.g., do these measures predict an explicit report of racism?).

The first set of comparisons relate to the hypothesized connection between ethnic attitudes and urban exposure. For this we will examine the relationship between individuals' ethnic effect on each task and:

[4] Others include face validity due to the nature of the task, the concurrent validity we obtain by comparing it to explicit reports of racism and other attitudes (which were collected in post-experiment surveys), and concurrent validity by comparing it to post-experiment reports about the legitimacy of the election results (as described in Branch 4, below).

- years living in nairobi (control for age)
- neighborhood diversity
- mixing of ethnicity in family (by parents or by marriage)
- rural ties index: first principal component of an index over questions regarding rural ties
- a second version of the rural ties that is compatible with the way it was measured in the rural areas (i.e. use only questions common to both rural and urban areas)[5]

A second set of comparisons checks whether these tests may have worked better among those who show more competence with the experimental setting, whether the behavior we record may be altered by knowledge of our interest in ethnicity, and whether the order in which the tasks were presented matters. We will compare behavior on each task to:

- performance on the literacy and Raven's test
- debrief – whether in their open-ended responses to a post-survey participants indicated that they thought the experiment was about ethnicity
- having heard about the experiment before taking it
- prime ordering – whether the coethnic partner (group) was presented first in the DG, PG, CYD (or second)

A third group of comparisons will be to compare these measures to explicit measures for purposes of seeing whether they predict ethnic attitudes as reported.

- support segregation
- fear walking down the street (2 questions)
- general fearfulness
- trust of people from the non-coethnic region of the country

For all of the above groups of comparisons, we will test correlations with confidence intervals, and note both the standard p-value and a Bonferroni corrected p-value across all comparisons.

For the comparisons to the explicit report (the final group above) we will also construct a single *explicit ethnic bias index* using PCA, and test its correlation to each task's ethnic effect measure. We will also test and report whether the relationship between the explicit

[5] The broader project of which this study is a part involved the collection of data from the AMP, DG, PG and CYD in rural as well as urban areas. The anslyses described in this pre-analysis plan, which involve data from the FA and WMT as well, are restricted to the urban sample.

report and our ethnic effect measures are related to responses on the two-item *effort to control racism* scale. We will do this simply by reporting separate correlations (between each task measure and the *explicit ethnic bias index*), for those who score 0, 1, or 2 on the *effort to control racism* scale. We expect positive correlations overall, but that they will be less positive for those who score higher on *effort to control racism.*

Branch 4: Follow-up Study

We were also able to conduct a follow-up study among a large portion of our original study participants, contacting them by phone to ask questions about their attitudes towards the Kenya Supreme Court's annullment of the country's recent election, which was held just after our data collection activities for the main project. The phone questionnaire contained 6 questions. Our primary interest is in understanding how measures taken during the lab experiments predict the responses to these questions, as a validation exercise.

Specifically we register three types of statistical tests. The first will take each response from the follow-up call, $election_1, ..., election_4, ..., election_6$, and examine its correlation with each of the laboratory based measures (each column of M_{econ} or M_{psych}). Confidence intervals and p-values will be obtained simply be regression. Results will be reported with and without a correction for multiple testing.

Second, rather than deal with 6 post-election measures, the first principal component will be constructed, and we will report the corresponding scree-plot, loadings, and share of variance explained. We will then check the correlation of this single dimension of post-election attitudes with the laboratory measures.

Third, while we had a very high recontact rate in the follow-up, it remains important to know how those who could not be reached might differ from those who could. We will thus examine differences in means on (a) background characteristics and (b) lab results (M_{econ} and M_{psych}), depending on whether people could or could not be reached. We do not enumerate all these characteristics here, as such a check is not central to our main purpose but rather just allows us to characterize how our sample may have changed between the original study and the follow-up survey.

Branch 5: Exploratory analyses

The above analyses characterize the main questions that motivate our study and the principal conclusions we hope to draw. There are many other comparisons we could make that either (a) we fail to anticipate; or (b) currently fall in the exploratory rather than confirmatory stage of our research cycle. This includes for example:

- comparisons by age, gender, education, sense of belonging, income, religion, news source (radio, tv, internet), etc.

- looking to the three questions we have about contact with the other ethnic group, and comparing those who report frequent "contact" with the other group with those who do not, or comparing those who further report this contact is “comfortable” with those who report differently.

To discipline our exploratory analyses and prevent overly optimistic inferences, we register our plan to use a split-sample approach for these additional exploratory analyses. Specifically, the procedure will be to:

1. *Split the data.* We will split the data into two equal-sized datasets, splitting on participant, blocking on session. We will conduct a series of balance tests to ensure that there are not chance differences on background covariates (without looking at any outcome data). If severe chance imbalances are found, a new split will be drawn. We will commit to a final split of the data prior to exploring any results involving outcomes. Once the data are split, one half will be labeled the *exploratory* set and the other the *confirmatory* set.

2. We will conduct all the registered anlayses described above in Branch 5 on the *exploratory* set first. Inspired by what we find, we will also conduct exploratory analyses at will. We will decided upon these exploratory anlayses and write fully functioning code to conduct them.

3. We will re-run all the registered analyse above on the *full* dataset (*exploratory* plus *confirmatory*) and report the results as our official results for the registered analysis.

4. We will run the exploratory analyses detailed in (2) above on the *confirmatory* set and report the results.

Any further analyses we run after the described procedure above (such as those suggested by colleagues or reviewers after we have looked at results on the confirmatory set) will be clearly marked in the text as exploratory-only without the benefit of the split sample. Thus we will have three types of analyses/results: (i) registered/ planned/ confirmatory (ii) exploratory analysis with split-sample validation, and (iii) purely exploratory. Our intention is to be extremely clear in describing what results belong in what category, though of course reviewers and editors may constrain our presentation in published work.